Humboldt-Universität zu Berlin

Institut für Bibliotheks- und informationswissenschaft

Berliner Handreichungen
zur Bibliotheks- und Informationswissenschaft

Heft 238

Moderne Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen

Projekte und Perspektiven

von

Alexandra Schneider

Moderne Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen

Projekte und Perspektiven

von
Alexandra Schneider

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 238

Schneider, Alexandra

Moderne Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen : Projekte und Perspektiven / von Alexandra Schneider. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2008. - 64 S. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 238)

Abstract:

Die vorliegende Arbeit beschäftigt sich mit modernen Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen. Wie die Verbindung der beiden gegensätzlich scheinenden Wortgruppen im Titel zeigt, werden in der Arbeit Aspekte aus der Informatik bzw. Informationswissenschaft mit Aspekten aus der Bibliothekstradition verknüpft. Nach einer kurzen Schilderung der Ausgangslage, der so genannten Informationsflut, im ersten Kapitel stellt das zweite Kapitel eine Einführung in die Theorie des Information Retrieval dar. Im Einzelnen geht es um die Grundlagen von Information Retrieval und Information-Retrieval-Systemen sowie um die verschiedenen Möglichkeiten der Informationserschließung. Hier werden Formal- und Sacherschließung, Indexierung und automatische Indexierung behandelt. Des Weiteren werden im Rahmen der Theorie des Information Retrieval unterschiedliche Information-Retrieval-Modelle und die Evaluation durch Retrievaltests vorgestellt. Nach der Theorie folgt im dritten Kapitel die Praxis des Information Retrieval. Es werden die organisationsinterne Anwendung, die Anwendung im Informations- und Dokumentationsbereich sowie die Anwendung im Bibliotheksbereich unterschieden. Die organisationsinterne Anwendung wird durch das Beispiel der Datenbank KURS zur Aus- und Weiterbildung veranschaulicht. Die Anwendung im Bibliotheksbereich bezieht sich in erster Linie auf den OPAC als Kompromiss zwischen bibliothekarischer Indexierung und Endnutzeranforderungen und auf seine Anreicherung (sog. Catalogue Enrichment), um das Retrieval zu verbessern. Der Bibliotheksbereich wird ausführlicher behandelt, indem ein Rückblick auf abgeschlossene Projekte zu Informations- und Indexierungssystemen aus den Neunziger Jahren (OSIRIS, MILOS I und II, KASCADE) sowie ein Einblick in aktuelle Projekte gegeben werden. In den beiden folgenden Kapiteln wird je ein aktuelles Projekt zur Verbesserung des Retrievals durch Kataloganreicherung, automatische Erschließung und fortschrittliche Retrievalverfahren präsentiert: das Suchportal dandelon.com und das 180T-Projekt des Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen. Hierbei werden jeweils Projektziel, Projektpartner, Projektorganisation, Projektverlauf und die verwendete

Technologie vorgestellt. Die Projekte unterscheiden sich insofern, dass in dem einen Fall eine große Verbundzentrale die Projektkoordination übernimmt, im anderen Fall jede einzelne teilnehmende Bibliothek selbst für die Durchführung verantwortlich ist. Im sechsten und letzten Kapitel geht es um das Fazit und die Perspektiven. Es werden sowohl die beiden beschriebenen Projekte bewertet als auch ein Ausblick auf Entwicklungen bezüglich des Bibliothekskatalogs gegeben.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin.

Online-Version: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h238/

Inhaltsverzeichnis

1 Einleitung

1.1 Gegenstand und Aufbau der Arbeit

Die vorliegende Arbeit beschäftigt sich mit modernen Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen. Unter Retrievalverfahren werden Verfahren der Repräsentation, Speicherung und Organisation von Information sowie der Zugriff auf Information, vor allem beim Online-Retrieval, verstanden.[1] Beim Information Retrieval geht es somit auf der einen Seite um die Gestaltung der Datengrundlage, z.B. durch Anreicherung oder automatische Indexierung, auf der anderen Seite um die Gestaltung der Suche durch gewisse Funktionalitäten der Sucheingabe sowie der Trefferausgabe, z.B. mittels Suchmaschinentechnologie. Im Mittelpunkt steht der „technisch-gestützte Prozess des Wissenstransfers vom Wissensproduzenten (klassisch: dem Autor) und dem Informations-Nachfragenden".[2] Auf den Bibliotheksbereich bezogen findet dieser Wissenstransfer bei der Erstellung und Benutzung des Bibliothekskatalogs, der klassischen bibliotheksbezogenen Anwendung schlechthin, statt. In heutiger Zeit handelt es sich bei dem Katalog um einen Online-Katalog, einen so genannten OPAC (Online Public Access Catalogue).

Wie die Verbindung der beiden gegensätzlich scheinenden Wortgruppen im Titel („moderne Retrievalverfahren" und „klassische bibliotheksbezogene Anwendungen") zeigt, werden in der vorliegenden Arbeit Aspekte aus der Informatik bzw. Informationswissenschaft mit Aspekten aus der Bibliothekstradition verknüpft. Durch die beispielhafte Darstellung des Einflusses von Methoden aus der Informatik auf das Bibliothekswesen bildet diese Arbeit eine Schnittmenge aus Informatik, Informationswissenschaft und Bibliothekswissenschaft, was sich in der verwendeten Terminologie niederschlägt. Die Terminologie der verschiedenen Wissenschaften ist uneinheitlich, ein Problem, das sich nicht in einer Arbeit von dem gegebenen Umfang lösen lässt. Informatik ist die „Wissenschaft, die sich mit den theoretischen Grundlagen, den Mitteln und Methoden sowie mit der Anwendung der Elektronischen Datenverarbeitung (EDV) beschäftigt, d.h. mit allen Aspekten der Informationsverarbeitung unter Einsatz von Computern einschließlich ihres Einflusses auf die Gesellschaft".[3] Unter Informationsverarbeitung wird die Erfassung und Übermittlung, Aufbereitung und Auswertung, Speicherung und Wiedergewinnung von Information verstanden. Im Rahmen der angewandten Informatik wurden sowohl Datenbanksysteme, die der Beschreibung, Speicherung und Wiedergewinnung von Information dienen, als auch Informationssysteme, die auf Datenbanken zurückgreifen, deren Information miteinander verknüpft und ausgewertet wird, entwickelt. Bei diesen beiden Systemen handelt es sich um allgemeine Anwendungen der Informatik; ein spezielles Anwendungsgebiet ist z.B. die Computerlinguistik, die bei der automatischen Indexierung eine Rolle spielt. Für die Erörterung der genannten Anwendungen werden die Fachbegriffe aus der Informatik verwendet.

Informationswissenschaft als Bindeglied zwischen Informatik und Bibliothekswissenschaft „beschäftigt sich mit Informations- und Kommunikationsprozessen in Wissenschaft, Wirtschaft und Verwaltung sowie zunehmend auch mit Fragen, die aus den Entwicklungen der In-

1 Vgl. G. Salton/M.J. McGill: *Information Retrieval – Grundlegendes für Informationswissenschaftler*. Hamburg: McGraw-Hill, 1987. S. 1.

2 Gerhard Knorz: „Information Retrieval-Anwendungen". In: M.G. Zilahi-Szabo (Hrsg.): *Kleines Lexikon der Informatik und Wirtschaftsinformatik*. München: Oldenbourg, 1995. S. 244.

3 Uwe Schneider/Dieter Werner (Hrsg.): *Taschenbuch der Informatik*, 5. Aufl. München und Wien: Carl Hanser Verlag, 2004. S. 25.

formationsgesellschaft resultieren".[4] Bibliothekswissenschaft im weiteren Sinn stellt den „systematisch geordneten Inbegriff aller wissenschaftlichen und technischen Erfahrungen auf dem Gebiet des Bibliothekswesens dar. Ihre Aufgabe ist die Erfassung und Analyse von Entwicklungen im Bereich der Informationsdistribution sowie auf dieser Grundlage die Entwicklung von Methoden und Theorien zur Informationsversorgung (hauptsächlich in der Wissenschaft)."[5] Im traditionellen Bibliothekswesen wird eine Terminologie verwendet, die sich stark von der der Informatik unterscheidet.

In der Informationswelt (Datenbanksysteme, Informationssysteme, Internet) haben sich die Retrievalsysteme weiterentwickelt, z.B. zu leistungsfähigen Suchmaschinen, was in der klassischen bibliothekarischen Welt – den Bibliothekskatalogen – zunächst wenig Entsprechung gefunden hat. Seit einigen Jahren hält diese Entwicklung aber auch in der Bibliothekswelt Einzug, wie die vorliegende Arbeit anhand verschiedener Beispiele zeigt.

Nach einer kurzen Schilderung der Ausgangslage, der so genannten Informationsflut, im ersten Kapitel stellt das zweite Kapitel eine Einführung in die Theorie des Information Retrieval dar. Im Einzelnen geht es um die Grundlagen von Information Retrieval und Information-Retrieval-Systemen sowie um die verschiedenen Möglichkeiten der Informationserschließung. Hier werden Formal- und Sacherschließung, Indexierung und automatische Indexierung behandelt. Des Weiteren werden im Rahmen der Theorie des Information Retrieval unterschiedliche Information-Retrieval-Modelle und die Evaluation durch Retrievaltests vorgestellt.

Nach der Theorie folgt im dritten Kapitel die Praxis des Information Retrieval. Es werden die organisationsinterne Anwendung, die Anwendung im Informations- und Dokumentationsbereich sowie die Anwendung im Bibliotheksbereich unterschieden. Die organisationsinterne Anwendung wird durch das Beispiel der Datenbank KURS zur Aus- und Weiterbildung veranschaulicht. Die Anwendung im Bibliotheksbereich bezieht sich in erster Linie auf den OPAC als Kompromiss zwischen bibliothekarischer Indexierung und Endnutzeranforderungen und auf seine Anreicherung (sog. Catalogue Enrichment), um das Retrieval zu verbessern. Der Bibliotheksbereich wird ausführlicher behandelt, indem ein Rückblick auf abgeschlossene Projekte zu Informations- und Indexierungssystemen aus den Neunziger Jahren (OSIRIS, MILOS I und II, KASCADE) sowie ein Einblick in aktuelle Projekte gegeben werden.

In den beiden folgenden Kapiteln wird je ein aktuelles Projekt zur Verbesserung des Retrievals durch Kataloganreicherung, automatische Erschließung und fortschrittliche Retrievalverfahren präsentiert: das Suchportal dandelon.com und das 180T-Projekt des Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen. Hierbei werden jeweils Projektziel, Projektpartner, Projektorganisation, Projektverlauf und die verwendete Technologie vorgestellt. Die Projekte unterscheiden sich insofern, dass in dem einen Fall eine große Verbundzentrale die Projektkoordination übernimmt, im anderen Fall jede einzelne teilnehmende Bibliothek selbst für die Durchführung verantwortlich ist.

Im sechsten und letzten Kapitel geht es um das Fazit und die Perspektiven. Es werden sowohl die beiden beschriebenen Projekte bewertet als auch ein Ausblick auf Entwicklungen bezüglich des Bibliothekskatalogs gegeben.

4 http://www.uni-saarland.de/fak5/fr56/ [Letzter Aufruf: 01.05.2007]
5 http://de.wikipedia.org/wiki/Bibliothekswissenschaft [Letzter Aufruf: 01.05.2007]

1.2 Ausgangslage

Seit drei Jahrhunderten liegt die Verdopplungsrate der Literatur bei etwa 15-20 Jahren. Die Zahl der publizierenden Autoren verdoppelt sich in gleicher Weise alle 15-20 Jahre.[6] Die große Zahl an Wissenschaftlern und die hohe Publikationsaktivität haben zu einer „Flut“ an wissenschaftlichen Publikationen, der sog. Literatur- oder Informationsflut geführt. Zu Büchern kommen graue Literatur und elektronische Medien hinzu. Gaus sieht vor allem folgende Ursachen der Informationsflut:

- es müssen nicht nur die heute geschaffenen Erkenntnisse, sondern auch das wichtigste früher geschaffene Wissen verfügbar sein, was zu einer Anhäufung wissenschaftlicher Erkenntnisse führt
- aufgrund der Schnelllebigkeit nimmt die Geltungsdauer vieler Informationen ab
- alle Arbeits- und Lebensbereiche (nicht nur die Wissenschaften) sind stark differenziert und spezialisiert geworden
- zwischen den Spezialgebieten ergeben sich viele Kombinationsmöglichkeiten und Wechselwirkungen; die interdisziplinäre Forschung nimmt zu
- Arbeit, Entwicklung und Forschung erfolgen in viel stärkerem Maße kooperativ; dies setzt gegenseitige Information voraus
- die Verbesserung der Kommunikationsmittel ermöglicht einen intensiven, weltweiten Informationsaustausch
- der Anteil der geistig Schaffenden nimmt zu, weshalb sowohl ein hoher Informationsbedarf als auch eine hohe Informationsproduktion entstehen[7]

Für das Bibliothekswesen bedeutet eine solche Entwicklung eine wachsende Zahl an Benutzern mit einem immer höheren Bedarf an Information. Aspekte der Beurteilung und Sicherung von Informationsqualität gewinnen an Bedeutung. Die Nutzer müssen rationell arbeiten können; Informations- und Wissensmanagement sind gefragt. Informationsmanagement kostet Geld, aber das vergebliche Suchen von Literatur, das Lesen von irrelevanter Literatur, unnötige Doppelforschung oder Doppelentwicklung als „Folgekosten des Nichtinformiertseins“[8] sind in der Regel mit weitaus höheren Kosten verbunden. Information wird ein immer wichtigerer Faktor in Gesellschaft und Wirtschaft.

Die Informationswelt befindet sich seit einigen Jahren in einem elementaren und anhaltenden Wandel. Mit der Entwicklung leistungsfähiger Rechner entstehen Möglichkeiten der Verarbeitung und Verwaltung auch größter Datenmengen. In der Folge sind immer mehr Informationen auf den unterschiedlichsten Wegen und in den unterschiedlichsten Medien zugänglich (u.a. in Universal- und Fachdatenbanken). Das Internet bietet eine kaum überschaubare und wenig strukturierte Flut an Daten. Durch Erstveröffentlichung von Literatur in elektronischer Form sowie Retrodigitalisierung geht die Tendenz zum vollständig maschinenlesbaren Text, dem sog. elektronischen Volltext. Parallel zu dem gesteigerten Informationsangebot haben sich auch die in Datenbanken eingesetzten Retrievalsysteme weiterentwickelt, ebenso wie für die Suche im Internet leistungsfähige Suchmaschinen entwickelt worden sind.

> Datenbanksysteme wie Suchmaschinen bieten fortschrittliche Methoden für das Retrieval in elektronischen Texten an, wie etwa eine natürlichsprachige Eingabe mit automatischer Fehlerkorrektur sowie eine nach Relevanz sortierte Ausgabe der Trefferdokumente (Relevance Ranking). Darüberhinaus wird dem Nutzer die Möglichkeit gegeben, aus den aus seiner Sicht

6 Vgl. Gisela Ewert und Walther Umstätter: *Lehrbuch der Bibliotheksverwaltung*. Stuttgart: Hiersemann, 1997. S. 1.

7 Wilhelm Gaus: *Dokumentations- und Ordnungslehre*, 5. Aufl. Berlin und Heidelberg: Springer, 2005. S. 23

8 Ebd., S. 26.

besonders relevanten Trefferdokumenten Suchbegriffe als Ausgangspunkt der weiteren Suche auszuwählen bzw. Suchbegriffe zu eliminieren, die in nicht relevanten Nachweisen aufgetreten sind (Relevance Feedback).[9]

Die geschilderte Entwicklung in der Informationswelt hält erst langsam in der klassischen bibliothekarischen Welt Einzug. Ein Beispiel ist der Einsatz von Verfahren der automatischen Indexierung in Bibliotheken. Anhänger der automatischen Indexierungsverfahren verweisen ebenfalls auf die Informationsflut, die die intellektuelle Behandlung heutiger Dokumentenmengen unmöglich macht, und versuchen darüber hinaus durch empirische Untersuchungen (Retrievaltests) zu belegen, dass automatische Verfahren, verglichen mit der intellektuellen Indexierung, mindestens gleich gute Resultate bei der Rückgewinnung von Dokumenten erzielen.[10]

Da im Rahmen der Schilderung der Ausgangslage bereits einige Fachtermini und Methoden wie automatische Indexierung, Retrievaltest, Relevance Ranking oder Relevance Feedback angesprochen wurden, werden sie im folgenden Kapitel, der Theorie des Information Retrieval, nun systematisch eingeführt und erläutert.

9 Hartmut Lohmann: *KASCADE: Dokumentanreicherung und automatische Inhaltserschließung.* Düsseldorf: Universitäts- und Landesbibliothek, 2000. S. 10.

10 Vgl. Holger Nohr: *Grundlagen der automatischen Indexierun*g, 3. Aufl. Berlin: Logos, 2005. S. 29.

2 Theorie des Information Retrieval

2.1 Grundlagen

2.1.1 Information Retrieval

Die Bezeichnung Information Retrieval stammt aus dem Englischen und lässt sich wörtlich übersetzen als Wiedergewinnung von Information. Hier kommt die Vorstellung zum Ausdruck, dass Information in großen Datenbeständen untergegangen ist und erst daraus wiedergewonnen werden muss. Im Glossar zu den *Grundlagen der praktischen Information und Dokumentation* wird Retrieval wie folgt erklärt:

> Retrieval (auch Recherche oder Information Retrieval genannt) bezeichnet den Arbeitsvorgang des gezielten Suchens bzw. Wiederfindens von relevanten Daten und Fakten zu einer speziellen Fragestellung in gedruckten oder elektronischen Informationsmitteln. Im heutigen Sprachgebrauch wird Recherche häufig mit dem Online-Retrieval gleichgesetzt. Bei der Online-Recherche werden Suchanfragen mit Hilfe der Retrievalsprache unter Verwendung von Operatoren formuliert und von einem Rechner im Direktzugriff auf eine Datenbank durchgeführt. Retrieval beschäftigt sich mit der Suche nach Informationen und mit der Repräsentation, Speicherung und Organisation von Wissen. Information Retrieval modelliert Informationsprozesse, in denen Benutzer aus einer großen Menge von Wissen die für ihre Problemstellung relevante Teilmenge suchen. Dabei entsteht Information, die im Gegensatz zum gespeicherten Wissen problembezogen und an den Kontext angepasst ist.[11]

In dem entsprechenden Artikel aus Wikipedia, der freien Online-Enzyklopädie, werden zwei Aspekte genannt, die das Information Retrieval prägen und es von der Suche in herkömmlichen Datenbanken abgrenzen: Vagheit und Unsicherheit:

> 1. Vagheit: Der Benutzer kann sein diffuses Informationsbedürfnis nicht präzise und formal (wie z.B. in SQL in relationalen Datenbanken) ausdrücken. Die Anfrage enthält daher vage Bedingungen.
> 2. Unsicherheit: Dem System fehlen Kenntnisse über den Inhalt der Dokumente (die Texte, Bilder, Video etc. enthalten können). Dies führt zu fehlerhaften und fehlenden Antworten. Probleme bei Texten bereiten z.B. Homographe [...] und Synonyme [...].[12]

Diese beiden Aspekte sind ebenfalls zentral für die Beschreibung der Aufgaben und Ziele der Fachgruppe „Information Retrieval“ innerhalb der Gesellschaft für Informatik:

> Die Fachgruppe „Information Retrieval“ in der Gesellschaft für Informatik beschäftigt sich dabei schwerpunktmäßig mit jenen Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen. Vage Anfragen sind dadurch gekennzeichnet, dass die Antwort a priori nicht eindeutig definiert ist. Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden können; häufig müssen zudem mehrere Datenbasen zur Beantwortung einer einzelnen Anfrage durchsucht werden. Die Darstellungsform des in einem IR-System gespeicherten Wissens ist im Prinzip nicht beschränkt (z.B. Texte, multimediale Dokumente, Fakten, Regeln, semantische Netze). Die Unsicherheit (oder die Unvollständigkeit) dieses Wissens resultiert meist aus der begrenzten Repräsentation von

11 Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 2: Glossar, 5. Aufl. München: Saur, 2004. S. 107.

12 http://de.wikipedia.org/wiki/Information_Retrieval [Letzter Aufruf: 01.05.2007]

dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten); darüber hinaus werden auch solche Anwendungen betrachtet, bei denen die gespeicherten Daten selbst unsicher oder unvollständig sind (wie z.B. bei vielen technisch-wissenschaftlichen Datensammlungen). Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte.[13]

Als kennzeichnend für das Gebiet des Information Retrieval werden somit vage Anfragen und unsicheres Wissen angesehen.

2.1.2 Information-Retrieval-Systeme

Generell sind am Information Retrieval zwei Personenkreise beteiligt, die Autoren und die Anwender mit Zielen und Aufgaben, wie die folgende Grafik veranschaulicht:

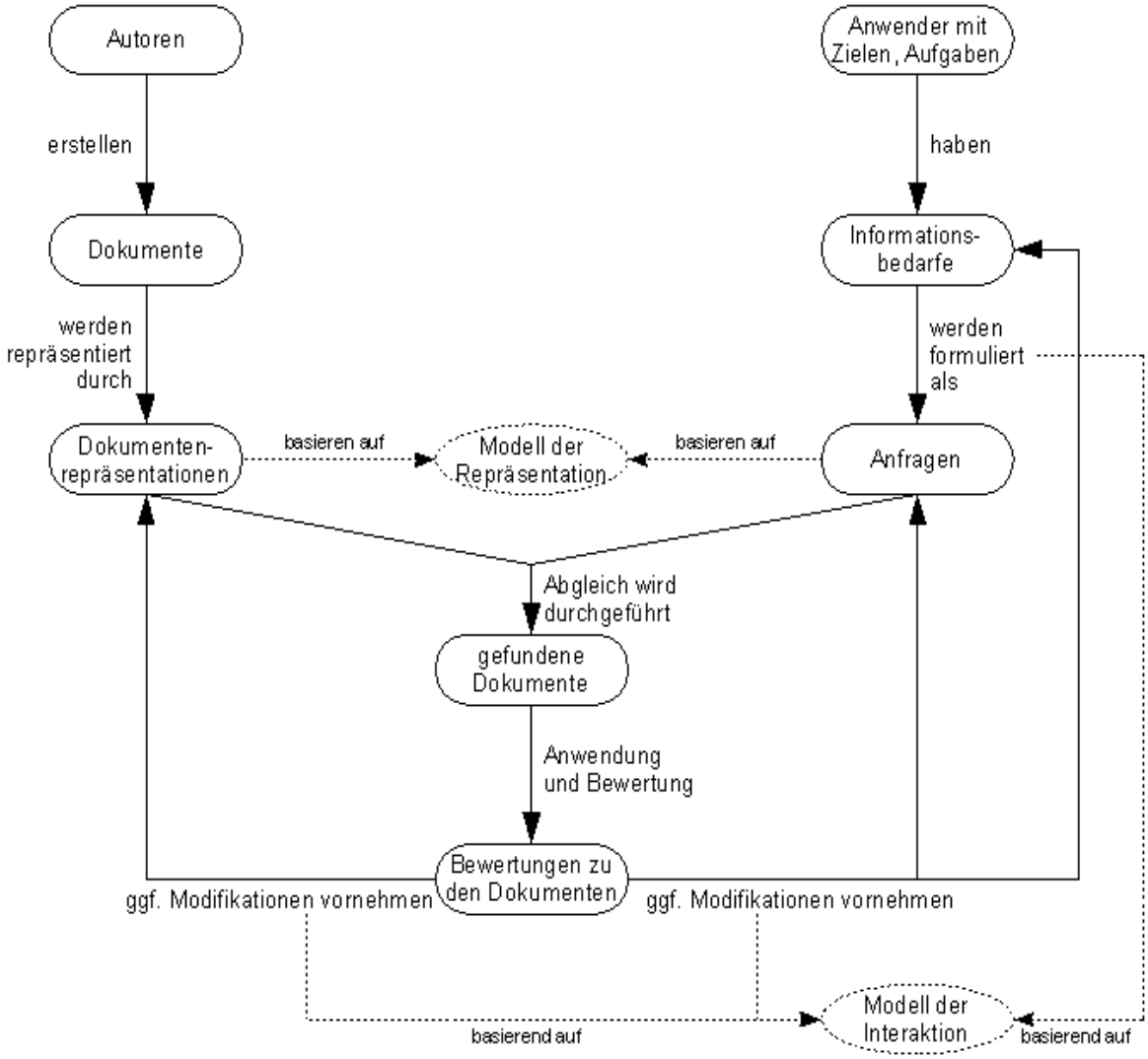

Abb. 1: Schematisches Modell des Information Retrieval nach Dominik Kuropka[14]

13 Zitiert nach: Norbert Fuhr: Information Retrieval. Skriptum zur Vorlesung im SS 06. S. 5-6. http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf [Letzter Aufruf: 01.05.2007]

14Quelle: http://de.wikipedia.org/wiki/Information_Retrieval [Letzter Aufruf: 01.05.2007]

Die Autoren stellen – aktiv oder passiv – die Dokumente einem Information-Retrieval-System zur Verfügung. Diese Dokumente werden vom System in eine für die Verarbeitung günstige Form (Dokumentenrepräsentation) umgewandelt. Die Anwender haben Ziele oder Aufgaben, für deren Lösung sie ihre Informationsbedarfe mit Hilfe von Anfragen decken müssen. Nach der Formulierung der Anfragen erfolgt im System der Abgleich zwischen Anfragen und eingestellten Dokumenten, der zur Ausgabe einer Ergebnis-/Trefferliste führt. Diese Liste muss vom Anwender beurteilt werden und begründet sein weiteres Vorgehen. Eventuell kommt es zu einer Änderung der Informationsbedarfe oder zu einer Verfeinerung der Anfrage oder zu einer Modifikation der Dokumentenrepräsentationen.

Die Form, in der die Informationsbedarfe formuliert werden müssen, und die Form der Interaktion hängt vom Modell des Information-Retrieval-Systems und seinen jeweiligen Modellen der Dokumentenrepräsentation sowie der Interaktion ab.

Ein Information-Retrieval-System ist eine Spezialisierung eines Informationssystems. Ein Informationssystem ist „eine Datenbank zusammen mit allen Programmen, die die Verarbeitung der in der Datenbank gespeicherten Informationen ermöglichen“.[15] „Information-Retrieval-Systeme (IRS) dienen der Informationsgewinnung aus Texten, multimedialen Dokumenten, Fakten usw. Sie operieren in der Regel auf unstrukturierten Daten.“[16] Nach Nohr ist es das Ziel von Information-Retrieval-Systemen, „wenig oder gänzlich unstrukturierte Informationen in einer Weise aufzubereiten, dass sie bei einem aktuellen Informationsbedürfnis mit entsprechenden Suchstrategien und -techniken möglichst präzise und vollständig wiederaufgefunden werden können“.[17] Information-Retrieval-Lösungen müssen somit ganzheitlich sein und „aufeinander abgestimmte Verfahren der Indexierung (der Repräsentation von Dokumenteninhalten über Metainformationen), der Speicherorganisation und der Recherche (des Zugriffs auf die erzeugten Inhaltsrepräsentationen) nach Informationen anstreben“.[18] Die Hauptelemente eines Information-Retrieval-Systems sind Input, Speicherung und Output. Zuerst müssen Informationen (z.B. Dokumente) auf elektronischem Wege in die Datenbank eingespeist werden (Input). Entweder liegen die Dokumente bereits digital vor oder sie müssen digitalisiert werden. Gedruckte Dokumente werden digitalisiert, indem sie eingescannt werden, woraufhin das durch den Scanvorgang erzeugte Bild mit Hilfe der optischen Zeichenerkennung (OCR) wieder in Text umgewandelt wird. Nach dem Input erfolgt die Speicherung der Dokumente, die aus der Ablage in einem Dateisystem bei gleichzeitiger Erzeugung von Metadaten durch Indexierung und Klassifizierung besteht. Der Output beinhaltet das Retrieval (Recherche in den Metadaten) und die Ausgabe der Dateien. Das gesamte System ist in der Regel ins Internet eingebunden (web-basiert).

Nach Gaus ist ein „ordentliches“ Retrievalsystem „zuverlässig, hat gute Dialog- und Eingabemöglichkeiten am Bildschirm und kann das Rechercheergebnis konfektionieren und benutzerfreundlich ausdrucken oder als E-Mail versenden“.[19] Der Bildschirmdialog für die Recherche sollte in zwei Fassungen vorliegen, für professionelle und für weniger geübte Rechercheure (strukturierte Rechercheformulare). Für die Korrektur bzw. Modifikation der Suchanfrage sollten bequeme Editoren zur Verfügung stehen. Nach Abschicken der Suchanfrage sollte zuerst die Anzahl der Treffer ausgegeben werden, dann die Titel und ggf. Abstracts bzw. Volltexte. Die Präsentation der selektierten Dokumentationseinheiten sollte benutzerfreundlich sein ebenso wie die Weiterverarbeitung der Ergebnisse (z.B. Downloading). Weite-

15 Uwe Schneider/Dieter Werner (Hrsg.): *Taschenbuch der Informatik*, a.a.O., S. 419.
16 Ebd., S. 452.
17 Holger Nohr: *Grundlagen der automatischen Indexierun*g, a.a.O., S. 20-21.
18 Ebd., S. 21.
19 Wilhelm Gaus: *Dokumentations- und Ordnungslehre*, a.a.O., S. 248.

re Anforderungen an das System sind sichere Speicherung, kurze Antwortzeiten, Zugreifbarkeit, einfache Handhabung und die Erfüllung der Aufgaben, wozu auch das Führen von Benutzungs- und Betriebsstatistiken gehört.

Im Weiteren wird unter einem Information-Retrieval-System die Gesamtheit verstanden, bei der neben die Informationswiedergewinnung die Informationserschließung tritt:

> Erst eine konzeptionell (und praktisch) gelöste Informationserschließung (Indexierung im weiteren Sinne) bildet nämlich die notwendige Voraussetzung für die Wiedergewinnung (Retrieval).[20]

Deshalb wird im Folgenden näher auf die Informationserschließung eingegangen, wobei der Schwerpunkt auf die Indexierung und hier wiederum auf die automatische Indexierung gelegt wird.

2.2 Informationserschließung

2.2.1 Formal- und Sacherschließung

Die Repräsentation von Dokumenten ist eine der Hauptaufgaben der Informationsversorgung durch Bibliotheken. Um Dokumente zu beschreiben und sie für das Retrieval such- und findbar zu machen, werden Metadaten verwendet:

> Metadaten sind Suchbegriffe, die den Dokumenten vom Produzenten beigegeben werden oder in einem bibliothekarischen Erschließungsvorgang intellektuell oder maschinell erzeugt werden. Traditionell gibt es in Büchern Metadaten in Form von Titelangaben (auf dem Titelblatt oder an einer sonstigen Titelstelle), Inhaltsverzeichnissen, Klappentexten oder Registern. Besondere Formen von Metadaten mit teilweise wertendem Charakter sind Rezensionen, Abstracts oder Bibliographien. Aus Sicht der bibliothekarischen Tradition sind Metadaten das Ergebnis der Formal- und der Sacherschließung; dazu gehören vor allem Personennamen, Körperschaftsnamen, Sachtitel, Erscheinungsorte, Erscheinungsjahre, Namen von Verlegern und Druckern, Schlagwörter, Gattungsbegriffe, Notationen.[21]

Die Bestandserschließung wird somit unterschieden in die Formal- und in die Sacherschließung. Bei der Formalerschließung handelt es sich traditionell um die formale Katalogisierung im alphabetischen Katalog, die nach bestimmten Regeln für die alphabetische Katalogisierung (RAK) erfolgt. Wesentliche formale Elemente sind hierbei der Sachtitel und der Verfassername. Als formale Erschließungsdaten gelten auch Identifikationsnummern wie ISBN (International Standard Book Number) oder auch die URL (Uniform Resource Locator) bei Netzpublikationen.[22]

Der Gegenstand der Sacherschließung hingegen ist nicht die formale, sondern die inhaltliche Beschreibung und Erschließung von Literatur. Die Sacherschließung lässt sich untergliedern in die verbale Sacherschließung und in die klassifikatorische Sacherschließung. Die verbale Sacherschließung verwendet hauptsächlich natürlichsprachige Bezeichnungen wie Schlagwörter und Stichwörter, während die klassifikatorische Sacherschließung vorwiegend auf Klassifikationssystemen mit hierarchisch geordneten Systemstellen (Notationen) beruht. Bei-

20 Jiri Panyr: *Automatische Klassifikation und Information Retrieval*. Tübingen: Niemeyer, 1986. S. 16.

21 Klaus Haller, Claudia Fabian: „Bestandserschließung“. In: Rudolf Frankenberger und Klaus Haller (Hrsg.): *Die moderne Bibliothek*. München: Saur, 2004. S. 223.

22 Für nähere Informationen zur Formalerschließung vgl. bibliothekarische Grundlagenwerke wie *Die moderne Bibliothek,* a.a.O., oder Rupert Hacker: *Bibliothekarisches Grundwissen*, 7. Aufl. München: Saur, 2000.

spiele für Klassifikationen sind die Dezimalklassifikation (DK), die Dewey Decimal Classification (DDC) oder die Allgemeine Systematik für Öffentliche Bibliotheken (ASB). Wie für die formale Erschließung gibt es auch für die verbale Sacherschließung durch Schlagwörter Regeln und Hilfsmittel: Regeln für den Schlagwortkatalog (RSWK) und Schlagwortnormdatei (SWD).

Die Tätigkeit der verbalen Sacherschließung wird auch als Indexieren, die der klassifikatorischen Sacherschließung als Klassieren bezeichnet. Der folgende Abschnitt führt in das Thema „Indexierung" ein.

2.2.2 Indexierung

Als Indexierung bezeichnet man „Verfahren, Methoden und Prinzipien der Inhaltserschließung von Texten (Dokumentarische Bezugseinheit) durch Zuweisung von inhaltskennzeichnenden Wörtern, den so genannten Index-Termini".[23] Nach Knorz drückt Indexieren „den Inhalt eines Dokumentes mit den Mitteln einer Dokumentationssprache (Vokabular und Syntax) aus".[24] Die Index-Termini heißen Stichwörter, wenn sie den Texten direkt entnommen werden, Schlagwörter, wenn sie einer Schlagwortliste entnommen sind, und Deskriptor, wenn sie einem geordneten und strukturierten Vokabular (z.B. Thesaurus) entstammen.[25]

Man kann verschiedene Formen der Indexierung unterscheiden:

- Extraktion und Addition (das Indexierungsverfahren betreffend)
- freie und kontrollierte Indexierung (das Indexierungssprachen-Vokabular betreffend)
- gleichordnende und syntaktische Indexierung (die Indexierungssprachen-Syntax betreffend)
- intellektuelle, computerunterstützte und automatische Indexierung (die Methode betreffend)[26]

Extraktionsverfahren übernehmen Termini aus dem Dokument (evtl. in bearbeiteter Form). Diese Stichwortindexierung wird vor allem von Verfahren der automatischen Indexierung angewendet. Additionsverfahren nehmen eine Repräsentation des Inhalts von Dokumenten durch sprachliche Elemente einer Indexierungssprache vor. Diese Elemente können aus dem Dokument stammen, müssen es aber nicht, sondern werden - in der Regel intellektuell – zugeteilt.

Beim Vokabular einer Indexierungssprache ist ein kontrolliertes, verbindliches Vokabular von einem offenen, freien zu unterscheiden. Bei der freien Indexierung werden beliebige, nicht vorgegebene Index-Termini verwendet, was zu einer ungenauen Verschlagwortung und zu Problemen bei der Recherche führt. Je nach Erschließungstiefe wird mit weiten oder engen Schlagwörtern gearbeitet. Im Rahmen der kontrollierten Indexierung werden Schlagwortkataloge oder Thesauri als kontrollierende und vereinheitlichende Instrumente eingesetzt, was

---

23 Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 2: Glossar, a.a.O., S. 52.

24 Gerhard Knorz: „Informationsaufbereitung II: Indexieren". In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 180-181.

25 Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 2: Glossar, a.a.O., S. 52.

26 Eine neuere Entwicklung auf dem Gebiet der Indexierung stellt das gemeinschaftliche Indexieren mit Hilfe von so genannter sozialer Software durch die Benutzer selbst dar. In diesem Zusammenhang spricht man von Tagging und von Tags.

vorteilhaft für das Retrieval ist. Dafür sind Erstellung und Pflege des Kontrollinstruments aufwändig. Probleme können auftreten, wenn das Vokabular der dokumentarischen Bezugseinheit aktueller ist als das des kontrollierten Vokabulars und sich deshalb kein passender Deskriptor finden lässt.

Bei der gleichordnenden Indexierung fehlt jegliche Syntax; die Index-Termini werden unabhängig von ihren inhaltlichen Zusammenhängen gleichrangig zugeordnet. Dagegen wird bei der syntaktischen (oder strukturierten) Indexierung mit einer Gewichtung der Index-Termini gearbeitet, z.B. einer Reihung nach Wichtigkeit, einer Klammerstruktur oder einer Nummer als Verbindungsdeskriptor.

Die DIN 31623-1 unterscheidet drei Indexierungsmethoden: intellektuelle, computerunterstützte und automatische Indexierung.[27] Der intellektuellen oder manuellen Indexierung geht eine intellektuelle Inhaltsanalyse voraus. Die Zuteilung der Index-Termini erfolgt durch den Menschen. Bei der computerunterstützten oder semiautomatischen Indexierung werden maschinell Index-Termini vorgeschlagen, die vom menschlichen Bearbeiter (Indexierer) angenommen, abgelehnt oder überarbeitet werden. Automatische oder maschinelle Indexierung durch den Computer benutzt im Wesentlichen statistische Textinformation sowie linguistische Methoden. Die Grenzen zwischen computergestützter und automatischer Indexierung sind fließend, denn auch bei so genannten automatischen Indexierungsverfahren fällt oft eine intellektuelle Nachbearbeitung (z.B. Wörterbuchpflege) an. Nach Mittelbach/Probst kann „bei keinem gegenwärtig im Einsatz befindlichen Verfahren von 'vollautomatischer Indexierung' gesprochen werden, da intellektuelle Analyse- und Kontrollschritte [...] immer zum *workflow* gehören“.[28] Die technischen Grundlagen der automatischen Indexierung werden im nächsten Abschnitt behandelt.

2.2.3 Automatische Indexierung

Nohr definiert automatische Indexierung als Verfahren, „die vollautomatisch Dokumente analysieren und abgeleitet aus dieser Analyse entweder ausgewählte Terme aus dem Dokument extrahieren und – unter bestimmten Verfahrensvoraussetzungen in einer bearbeiteten Form – als Indexterme abspeichern (Extraktionsverfahren) oder Deskriptoren einer kontrollierten Indexierungssprache dem Dokument als Inhaltsrepräsentanten zuweisen (Additionsverfahren)“.[29]

In der Literatur üblich ist eine Einteilung automatischer Indexierungsverfahren in computerlinguistische, statistische und begriffsorientierte Verfahren. Kurz vorangestellt sei noch der zeichenkettenorientierte Ansatz (einfache Stichwortextraktion/Volltextinvertierung): Die Verfahren dieses Ansatzes betrachten jeden Text isoliert für sich und fassen ihn als Folge von Wörtern (und jedes Wort als Folge von Buchstaben) auf. Alle Wörter (Zeichenketten) eines Textes, die nicht in einer Stoppwortliste enthalten sind, werden vollautomatisch in invertierten Listen abgelegt. Dieser Ansatz kann aber nicht wirklich zu den Indexierungsverfahren gezählt werden, weil weder eine Auswahl getroffen wird noch Wörter bearbeitet werden. Der Aufwand wird auf den Suchenden verlagert.

---

27 Vgl. DIN 31623-1: „Indexierung zur inhaltlichen Erschließung von Dokumenten“. In: *Publikation und Dokumentation 2*, 3. Aufl. Berlin: Beuth, 1989. S. 275-279.

28 Jens Mittelbach und Michaela Probst: *Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung. Strategien für das Bibliothekssystem der Freien Universität* Berlin. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2006 (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 183). S. 23-24.

29 Holger Nohr: *Grundlagen der automatischen Indexierun*g, a.a.O., S. 27.

Die Ausführungen zur Typologie echter automatischer Indexierungsverfahren folgen weitestgehend der Darstellung von Bertram.[30] Bertram listet vorab folgende typische Aufgaben auf, die automatische Verfahren zu bewältigen haben:

- Stoppwort-Eliminierung (Ausschaltung von nicht-bedeutungstragenden Wörtern)
- Wortformenreduktion (Zurückführung von grammatikalischen Flexionsformen auf Grund- oder Stammformen; auch Lemmatisierung oder Stemming genannt)
- Dekomposition (Zerlegung von Komposita in sinnvolle begriffliche Bestandteile)
- Wortgruppenerkennung (Erkennung von Wörtern mit gemeinsamem Sammelwort und Paraphrasen)
- Namenserkennung (Erkennung von Personen-, Orts-, Institutionennamen etc.)

Diese Aufgaben werden durch computerlinguistische Verfahren gelöst; Stoppwörter können aber auch über Worthäufigkeiten (statistische Verfahren) bestimmt werden. Generell problematisch für die automatische Indexierung sind implizite Inhalte, Paraphrasen, Umgangssprache, Metaphern, Homonyme, Polyseme und Neologismen.

Computerlinguistische Verfahren identifizieren Index-Termini auf Grundlage einer vorherigen linguistischen Analyse, die auf drei Ebenen der Sprache ansetzen kann:

- morphologische Analyse
- syntaktische Analyse (so genanntes Parsing)
- semantische Analyse

Morphologische Analysen finden auf der Wortebene, Syntaxanalysen auf der Wortgruppen-/Satzebene und semantische Analysen auf der Ebene des ganzen Dokuments statt. Computerlinguistische Verfahren sind abhängig vom gegebenen Sprachsystem, d.h. es bestehen z.B. große Unterschiede zwischen der deutschen und der englischen Sprache, da die englische Sprache flexionsärmer und morphologisch wenig komplex ist. Die Verfahren lassen sich nach der zur Anwendung kommenden Technik in regel- und wörterbuchbasierte Verfahren unterteilen. Regelbasierte Verfahren führen die linguistische Analyse auf Grundlage von Regeln durch, die in Form von Algorithmen formuliert werden. Hierbei handelt es sich um eine Verarbeitungsvorschrift, aus der die Abfolge der einzelnen Verarbeitungsschritte eindeutig hervorgeht. Die Regeln werden sequentiell durchlaufen, sodass die jeweils erste passende Regel Anwendung findet. Bei diesen Verfahren fällt der einmalige Aufwand zur Erstellung von Regeln, aber keine weitere Pflege an. Für flexions- und kompositumreiche Sprachen wie die deutsche sind regelbasierte Verfahren weniger geeignet. Wörterbuchbasierte Verfahren führen die linguistische Analyse aufgrund eines hinterlegten Wörterbuchs durch. Die Behandlung von Komposita lässt sich somit einfacher als durch Regeln realisieren. Dafür fällt bei diesen Verfahren der Aufwand der kontinuierlichen Wörterbuchpflege an. In der Praxis können Kombinationen beider Ansätze sinnvoll sein genauso wie Kombinationen von computerlinguistischen mit statistischen Verfahren.

Unter statistischen Verfahren fasst man Systeme zusammen, die auf der Termgewichtung beruhen. Sie vergleichen die Repräsentationen von Frage- und Dokumentinhalt miteinander und liefern als Antwort gewichtete Treffer. Hierbei spielen Häufigkeiten eine entscheidende Rolle (Termfrequenzansatz). Zum einen geht es um die Anzahl eines Terms im Dokument. Diese Zahl wird ins Verhältnis gesetzt zu der Gesamtzahl der Terme im Dokument, womit man die relative Termhäufigkeit erhält. Diese wiederum wird ins Verhältnis gesetzt zur Anzahl der Dokumente, in denen der Term vorkommt (inverse Dokumenthäufigkeit). Stoppwörter z.B. sind hochfrequent, aber zu wenig bedeutungstragend. Niedrigfrequente Wörter sind für den

30 Vgl. Jutta Bertram: *Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente*. Würzburg: Ergon, 2005. S. 97-114.

Fragenden zu wenig vorhersehbar. Deshalb sind Terme im mittleren Frequenzbereich am entscheidungsstärksten. Für eine sinnvolle Frequenz müssen intellektuell Schwellenwerte festgelegt werden, damit Stoppwörter oder andere zu wenig bedeutungstragende Wörter anhand von Worthäufigkeiten bestimmt werden können.

Begriffsorientierte oder wissensbasierte Verfahren versuchen sich von der benennungsorientierten Ebene dadurch zu lösen, dass sie die inhaltskennzeichnenden Terme aus einem zugrunde liegenden kontrollierten Vokabular zuteilen. Indem sie auf die Bedeutung der Inhalte schließen, simulieren sie das Arbeitsergebnis intellektueller Inhaltserschließung. Modelle aus der Künstlichen Intelligenz beziehen Weltwissen (nach Reimer „Hintergrundwissen über den Diskursbereich, aus dem die zu bearbeitenden Dokumente stammen“[31]) ein und erweitern ihre Wissensbasis selbst. Sie bauen das benötigte Vorwissen selbst auf und lernen. Als Wissensquelle können ein Lexikon oder Texte aus dem entsprechenden Dokumentenbestand dienen. Für die Erschließung der Bedeutung von Wörtern ist der jeweilige Kontext entscheidend. Analyseverfahren müssten daher den Kontext auftretender Wörter berücksichtigen, was bisherige Verfahren nicht oder nur unzureichend leisten. Eine Lösung wird von Mustererkennungsverfahren (Pattern-Matching-Verfahren) erhofft.[32]

Textanalyse und Indexierung stellen den Eingangsschritt des Information Retrieval dar. Die Anwendung automatischer Indexierungsverfahren ermöglicht erst bestimmte fortschrittliche Retrievalansätze. Das eigentliche Retrieval kann auf verschiedene Arten modelliert werden, die im folgenden Kapitel vorgestellt werden.

2.3 Information-Retrieval-Modelle

Indexierung und Recherche sind aufeinander abzustimmende Teilprozesse, die deshalb einem gemeinsamen Modell des Information Retrieval unterworfen sein müssen, das Indexierungs- und Retrievalprozess gleichermaßen berücksichtigt.

Im Glossar zu den *Grundlagen der praktischen Information und Dokumentation* wird Information-Retrieval-Modell wie folgt erklärt:

> Information-Retrieval-Modelle spezifizieren, wie zu einer Anfrage die Antwortdokumente aus einer Dokumentensammlung bestimmt werden. Dabei macht jedes Modell bestimmte Annahmen über die Struktur von Dokumenten und Anfragen und definiert daraus die sogenannte Retrievalfunktion, die das Retrievalgewicht (Gewichtung) eines Dokuments bezüglich einer Anfrage bestimmt. Die Dokumente werden dann nach fallenden Gewichten sortiert und dem Benutzer präsentiert. Boolesches Retrieval bestimmt eines der Gewichte 0 oder 1. Andere Ansätze arbeiten mit Wahrscheinlichkeiten, z.B. im Hinblick auf Relevanz (Probabilistisches Retrieval) oder verwendetes Vokabular (Sprachmodell), mit gewichteter Indexierung (Fuzzy-Retrieval) oder mit geometrischen Interpretationen (Vektorraum-Modell).[33]

Nohr unterscheidet Modelle des Information Retrieval grundsätzlich in Exact-Match-Modelle und Best-Match-Modelle. Exact-Match-Modelle (meistens Boolesches Retrieval) teilen eine

31 Ulrich Reimer: „Verfahren der automatischen Indexierung“. In: Rainer Kuhlen (Hrsg.): *Experimentelles und praktisches Information Retrieval*. Konstanz: Universitätsverlag, 1992. S. 180.

32 Nohr sieht die Mustererkennungsverfahren sogar als eigenständige Gruppe auf derselben Ebene wie statistische, informationslinguistische und begriffsorientierte Verfahren. Vgl. Holger Nohr: *Grundlagen der automatischen Indexierung*, a.a.O., S. 88-92.

33 Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 2: Glossar, a.a.O., S. 54.

Dokumentenkollektion in zwei Teile: Dokumente, die den „exact match“ erfüllen, und Dokumente, die ihn nicht erfüllen. Dokumente sind also relevant oder nicht; Abstufungen sind nicht möglich. Best-Match-Modelle (Vektorraum-Modelle, probabilistische Modelle) hingegen vergeben einen Grad der Relevanz für jedes Dokument.[34] Die durch eine Suchanfrage nachgewiesenen Dokumente werden in einer Rangfolge (Ranking) ausgegeben, die der Ähnlichkeit der Dokumente mit der Suchanfrage entspricht. Man spricht deshalb häufig von Verfahren des „Relevance Ranking“:

> Relevance Ranking erlaubt die Priorisierung der aufgefundenen Dokumente innerhalb einer Treffermenge. Der Priorisierung liegt das Konstrukt der Relevanz zugrunde, einer Ähnlichkeitsbeziehung, die zwischen einer Anfrage und einem Dokument besteht bzw. anhand zu definierender Ähnlichkeitsmerkmale hergestellt wird.[35]

Nach Fuhr sind allen Retrievalmodellen die folgenden grundlegenden Charakteristika gemein: Jedes Modell macht Annahmen über die Struktur von Dokumenten und Fragen. Ein Dokument kann entweder als Menge oder Multimenge von so genannten Termen aufgefasst werden, wobei im zweiten Fall das Mehrfachvorkommen berücksichtigt wird. Ein Term stellt einen Suchbegriff dar, der ein einzelnes Wort, einen mehrgliedrigen Begriff oder auch ein komplexes Freitextmuster umfassen kann. Die Menge aller Terme in der Dokumentenkollektion stellt das Indexierungsvokabular dar. In der Dokumentenrepräsentation können die Terme gewichtet sein, was Aufgabe der im vorigen Kapitel beschriebenen Indexierung ist. Man kann somit zwischen ungewichteter (Gewicht eines Terms ist entweder 0 oder1) und gewichteter Indexierung (das Gewicht ist eine nicht-negative reelle Zahl) unterscheiden. Ebenso wie bei Dokumenten können auch die Terme in der Frage entweder ungewichtet oder gewichtet sein. Daneben unterscheidet man zwischen linearen (Frage als Menge von Termen, ungewichtet oder gewichtet) und Booleschen Anfragen.[36]

Man differenziert in der Literatur üblicherweise die folgenden Modelle:

- mengentheoretische Modelle (Boolesches Retrieval, erweitertes Boolesches Retrieval und Fuzzy-Retrieval)
- algebraische Modelle (z.B. vektorraum-basierte Modelle)
- probabilistische Modelle (binäre Unabhängigkeit, unsichere Inferenz, statistische Sprachmodelle)

Mengentheoretische Modelle führen die Ähnlichkeitsbestimmung von Dokumenten auf die Anwendung von Mengenoperationen zurück. Beim Booleschen Retrieval sind die Frageterme ungewichtet und durch Boolesche Operatoren (AND, OR, NOT) miteinander verknüpft. Die Dokumente haben eine ungewichtete Indexierung, weshalb Boolesches Retrieval nur Retrievalgewichte von 0 und 1 liefert. Für Suchanfragen, die entsprechend der Booleschen Logik gebildet sind, werden alle relevanten Dokumente gefunden. Boolesches Retrieval hat Vor- und Nachteile:

> Zwar ist diese scharfe Trennung zwischen gefundenen und nicht gefundenen Dokumenten von Vorteil, wenn man nach formalen Kriterien sucht; bei inhaltsbezogenen Kriterien ist das Ignorieren von Vagheit und Unsicherheit durch die fehlende Gewichtung aber von Nachteil. Zudem

34 Vgl. Holger Nohr: Grundlagen der automatischen Indexierung, a.a.O., S. 138-139.

35 Ebd., S. 140.

36 Vgl. Norbert Fuhr: „Theorie des Information Retrieval I: Modelle“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 207.

sind die meisten Endnutzer mit der Verwendung der Booleschen Logik überfordert.[37]
Fuzzy Retrieval verwendet die gleiche Struktur der Anfragen, allerdings in Kombination mit gewichteter Indexierung. Im Vergleich zu anderen Verfahren liefern sowohl Fuzzy Retrieval als auch Boolesches Retrieval eine relativ schlechte Retrievalqualität und sind wenig benutzerfreundlich.[38]

Algebraische Modelle stellen Dokumente und Anfragen als Vektoren, Matrizen oder Tupel dar und berechnen paarweise Ähnlichkeiten. Dem Vektorraummodell liegt eine geometrische Interpretation zugrunde, bei der Dokumente und Anfragen als Punkte in einem Vektorraum aufgefasst werden, der durch die Terme der Kollektion aufgespannt wird. Anfragen besitzen somit eine lineare Struktur, wobei die Frageterme aber gewichtet sein können. Nach Fuhr haben zahlreiche experimentelle Untersuchungen die hohe Retrievalqualität des Vektorraummodells belegt. Die meisten Web-Suchmaschinen basieren auf diesem Modell.[39] Das Vektorraummodell ist eines der Modelle, die Relevanzrückkopplung (Relevance Feedback) ermöglichen. Hierbei werden dem Benutzer einige Antwortdokumente zu einer initialen Anfrage vorgelegt, die er bezüglich ihrer Relevanz beurteilen soll. Aus diesen Urteilen kann man dann eine modifizierte Frageformulierung berechnen, die in der Regel zu besseren Antworten führt. Durch die Relevanzrückkopplung ergibt sich eine deutliche Verbesserung der Retrievalqualität, aber bei Experimenten mit realen Benutzern stellte sich heraus, dass diese oft wenig bereit sind, Relevanzurteile abzugeben.[40]

Während die bisher beschriebenen Modelle unterschiedliche Arten von Ähnlichkeiten zwischen Frage- und Dokumentbeschreibungen berechnen, sehen probabilistische Modelle die Bestimmung von Dokumentenähnlichkeiten als ein mehrstufiges Zufallsexperiment an und greifen auf Wahrscheinlichkeiten und probabilistische Theoreme zurück. Sie schätzen die Relevanzwahrscheinlichkeit, dass das Dokument auf die Frage als relevant beurteilt wird, und ordnen die Dokumente nach dieser Wahrscheinlichkeit (probabilistisches Ranking-Prinzip). Eine solche Rangordnung führt zu optimaler Retrievalqualität.[41] Als Erweiterung der probabilistischen Modelle gibt es eine logische Sicht auf Information-Retrieval-Systeme. Es wird angenommen, dass man beim Retrieval nach Dokumenten sucht, die die Anfrage logisch implizieren. Steht ein Thesaurus zur Verfügung, so kann man die darin enthaltenen hierarchischen Beziehungen als logische Implikationen auffassen (z.B. Rodeln – Wintersport). Mit diesem zusätzlichen Wissen impliziert das Dokument auch die neue Anfrage. Um über Boolesches Retrieval hinauszugehen, muss man unsichere Inferenz zulassen. Auch wenn der Anfrageterm nicht direkt im Dokument vorkommt, so besteht dennoch eine gewisse Wahrscheinlichkeit, dass das Dokument auf die Anfrage relevant ist. Mit solchem unsicheren Wissen würde das Dokument wieder die Anfrage (unsicher) implizieren. Es besteht ein Zusammenhang zwischen der Implikationswahrscheinlichkeit und der Relevanzwahrscheinlichkeit.

Die beschriebenen Modelle lassen sich nach ihrem mathematischen Fundament und ihren Eigenschaften (ohne bzw. mit Terminterdependenzen) klassifizieren, wie die folgende Grafik veranschaulicht:

37 Ebd., S. 208.
38 Ebd.
39 Ebd., S. 209.
40 Ebd., S. 210.
41 Ebd., S. 211.

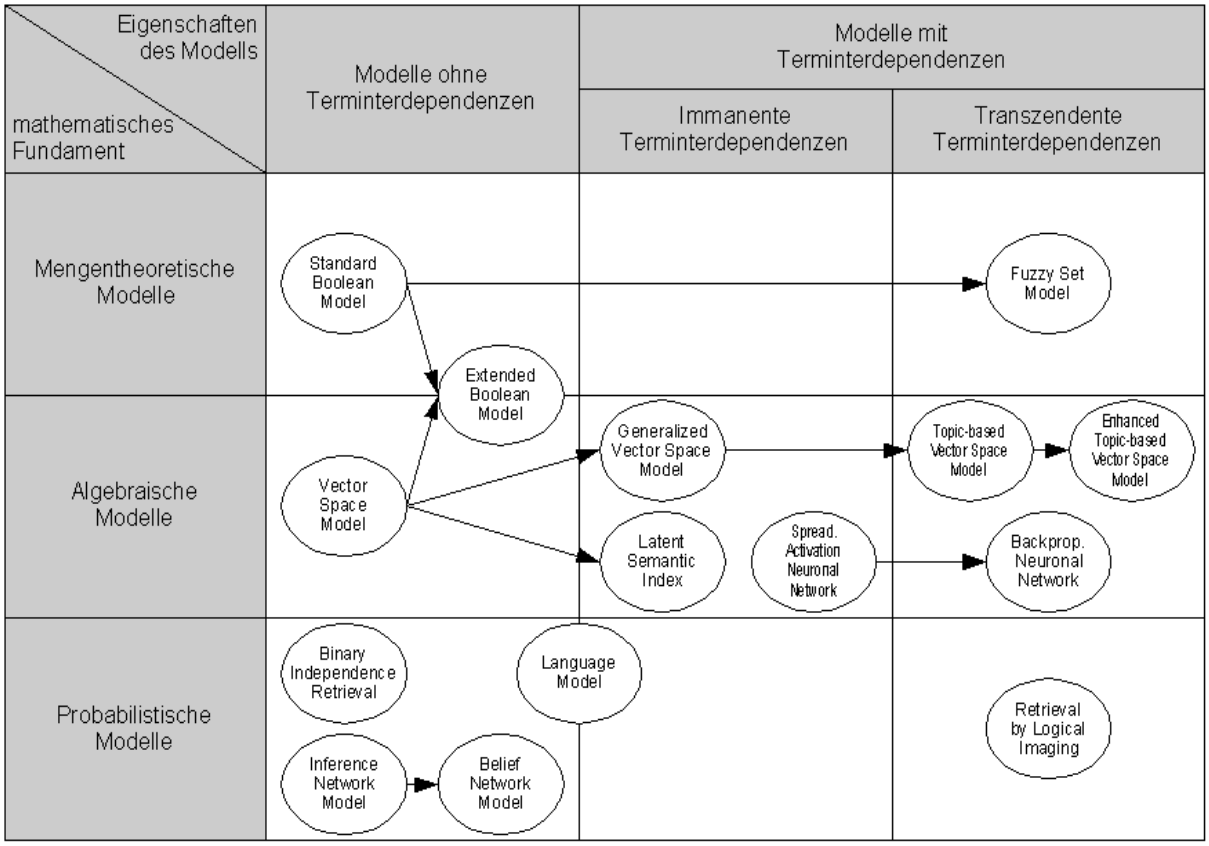

Abb. 2: Klassifikation von Information-Retrieval-Modellen nach Dominik Kuropka[42]

Nach dieser Übersicht über die gängigen Information-Retrieval-Modelle geht es im nächsten Kapitel um die Evaluation.

2.4 Evaluation durch Retrievaltests

Information-Retrieval-Systeme wurden bereits sehr früh unter dem Gesichtspunkt der Bewertung betrachtet. Bei der Bewertung unterscheidet man die Effizienzbewertung von der Effektivitätsbewertung. Die Effizienzbewertung oder auch Kosten-Nutzen-Analyse setzt wirtschaftliche Faktoren mit den Retrievalergebnissen in Relation (Kosten pro Dokument, Antwortzeitverhalten, maschinelle und menschliche Ressourcen u.Ä.); die Effektivitätsbewertung beschäftigt sich mit der Qualität der Antwortdokumente. In letzter Zeit kommt der Begriff der Performanz hinzu, der Effizienz und Effektivität verbindet, aber noch nicht standardisiert ist.[43]

Von Fragestellungen wie z.B. der Gestaltung der Benutzungsschnittstelle aus softwareergonomischer Sicht abgesehen gilt das Hauptinteresse der Evaluation der Indexierungsqualität. Nach Nohr geht es hierbei entweder um den Vergleich zwischen manueller und automatischer Indexierung oder um den Vergleich verschiedener Verfahren der automatischen Indexierung bzw. möglicher Verfahrenskombinationen oder um die Frage nach der Eignung bestimmter Indexierungssysteme für eine konkrete Anwendungssituation.[44]

42Quelle: http://de.wikipedia.org/wiki/Information_Retrieval [Letzter Aufruf: 01.05.2007]

43 Vgl. Christa Womser-Hacker: „Theorie des Information Retrieval III: Evaluierung“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, a.a.O., S. 227.

44 Vgl. Holger Nohr: *Grundlagen der automatischen Indexierung*, a.a.O., S. 149-150.

Im Rahmen der Effektivitätsbewertung muss die Effektivität, d.h. die Fähigkeit, relevante Dokumente wiederaufzufinden und zu liefern sowie gleichzeitig nicht-relevante Dokumente zurückzuhalten, gemessen werden. In so genannten Retrievaltests, in denen eine immer stärkere Einbeziehung der Benutzer stattfindet, wird der Retrievaloutput als Bemessungsgrundlage herangezogen. „Erst im Zeitalter der Suchmaschinen ist man zu der Ansicht gelangt, dass den Benutzern der Systeme eine entscheidende Rolle bei der Bewertung zukommt.“[45] Da es beim Retrievaloutput um relevante bzw. nicht-relevante Dokumente geht, ist der Begriff der Relevanz entscheidend. Bei Ranking-Systemen wie den gängigen Suchmaschinen spielt zugleich die Positionierung der gefundenen Dokumente eine wichtige Rolle. Die relevantesten Dokumente sollen auf den ersten Positionen der Ergebnisliste erscheinen, um den Benutzer zufrieden zu stellen. Die Relevanzbestimmung bildet somit die Basis der Effektivitätsbemessung, was als widersprüchlich empfunden wird:

> Häufig ist es jedoch gerade die Relevanzbestimmung, welche Kritik an der Retrievalmessung hervorruft. Es wird ein Widerspruch zwischen der statistisch-quantitativen Anwendung von Maßen und der relativ unscharfen, nur schwer in quantitativen Kategorien fassbaren Basis der Relevanzbewertung gesehen. Das traditionelle Verständnis des Relevanzbegriffs geht von einer Relation zwischen einer bestimmten Anfrage und den Ergebnisdokumenten aus. Die Forderung nach objektiver Relevanzbestimmung durch einen unabhängigen Juror ist schwer einlösbar. Einige wenige Untersuchungen analysieren die Relevanzurteile sowie die Umstände ihrer Abgabe, aber auch die Kategorie der „subjektiven Relevanz“, die durch verschiedene Benutzerbedürfnisse und Relevanzvorstellungen entstehen kann.[46]

Weitere Aspekte der subjektiven Relevanz sind das Vorverständnis des Fragenden und seine weiteren persönlichen Voraussetzungen (z.B. Verständnis einer Sprache oder Vertrautheit mit dem Gegenstand der Suchanfrage). Zum Teil wird deshalb in Tests der Begriff der Relevanz durch den der Nützlichkeit ersetzt. Nohr beschreibt zwei grundsätzliche Bezugsrahmen für den Relevanzbegriff und die Relevanzbeurteilung im Rahmen von Retrievaltests. Der erste Bezugsrahmen „Relevanzbeurteilung als Aneignung neuen Wissens“ ist völlig vom Informationsbedarf bzw. -defizit des Nutzers abhängig und somit subjektiv. Der zweite Bezugsrahmen wird als „vom gleichen Thema handelnd“ bezeichnet. Relevanz wird in diesem Fall an einer begrifflichen Übereinstimmung zwischen Dokument und Frage festgemacht, also als eine Eigenschaft von Dokument und Frage betrachtet (objektive Relevanzdefinition). Objektivität darf jedoch nicht mit Eindeutigkeit verwechselt werden, denn es gibt graduelle Abstufungen der Relevanz von Antwortdokumenten.[47]

Die Relevanzmessung basiert in der Regel auf den Standardmaßen Recall (Vollzähligkeitsrate) und Precision (Relevanzrate oder Präzision), weil sie am weitesten verbreitet, einfach zu interpretieren und ihre Schwachstellen bekannt sind. Der Recall misst die Vollständigkeit des Retrievalergebnisses, also wie viele relevante Dokumente in der Ergebnisliste enthalten sind. Er stellt das Verhältnis zwischen selektierten bzw. nachgewiesenen relevanten Dokumenten und im gesamten Dokumentenbestand vorhandenen relevanten Dokumenten dar. Sein Wertebereich liegt zwischen 0 und 1, wobei 0 das schlechteste und 1 das bestmögliche Ergebnis darstellt. Die Menge der relevanten, aber nicht nachgewiesenen Dokumente wird als Verlustrate bezeichnet. Bei umfangreichen Tests ist es unmöglich, die Anzahl aller im gesamten Dokumentenbestand vorhandenen relevanten Dokumente herauszufinden, weshalb für diese Zahl ein Schätzwert angenommen wird. Die Ermittlung des Schätzwerts erfolgt durch eine möglichst genaue Annäherung an die Gesamtzahl aller relevanten Dokumente. Hierfür kom-

45 Christa Womser-Hacker: „Theorie des Information Retrieval III: Evaluierung“, a.a.O., S. 227.
46 Ebd., S. 227-228.
47 Holger Nohr: *Grundlagen der automatischen Indexierung*, a.a.O., S. 152-153.

men als Methoden die Known-Item-Search (Suche nach einem bekannten Dokument), die Generalisierung auf der Basis eines genau bewerteten, repräsentativen Subset, die Schätzung durch Experten und die Pooling-Methode in Frage.[48]

Da der Recall die Ballastquote nicht einbezieht, wird als komplementäres Maß die Precision zur Messung der Genauigkeit eines Retrievalergebnisses herangezogen. Sie stellt das Verhältnis zwischen selektierten relevanten Dokumenten und allen selektierten Dokumenten dar. Ihr Wertebereich liegt ebenfalls zwischen 0 und 1. Die Menge der durch das Information-Retrieval-System nachgewiesenen, aber nicht-relevanten Dokumente wird als Noise, Rauschen oder Ballast bezeichnet.

Recall und Precision können durch die Indexierung beeinflusst werden. Die Ausweitung der Indexierungstiefe ist theoretisch mit einer Erhöhung des Recall bei gleichzeitiger Reduzierung der Precision verbunden. Umgekehrt nimmt man bei der Spezifierung der Indexierung eine Erhöhung der Precision zuungunsten des Recall an. Hier liegt also ein umgekehrtes Verhältnis vor, weshalb die gemeinsame Nutzung der beiden Maße sinnvoll erscheint. In der Praxis werden Recall und Precision durch die Suchstrategie der Benutzer beeinflusst, beispielsweise durch die Expansion der Suchanfrage. Im Zusammenhang mit dem Benutzer wurden so genannte Benutzerstandpunkte („elementary viewpoints“) eingeführt, indem Benutzer z.B. über folgende Stereotypen modelliert wurden:

- Dem Benutzer genügt ein relevantes Dokument.
- Der Benutzer möchte alle relevanten Dokumente zu einem Thema und ist bereit, Ballast in Kauf zu nehmen.
- Der Benutzer bricht nach fünf aufeinander folgenden irrelevanten Dokumenten ab.

Es ist fraglich, ob diese Stereotypen mit den Benutzervorstellungen in der Realität übereinstimmen. Aufgrund der unübersichtlichen Menge an Informationen liegt heute der Schwerpunkt auf der Filterung und Zurückhaltung des Ballasts, d.h. die Bedeutung der Precision nimmt zu.[49]

Bei der Durchführung von Retrievaltests konkurrieren zwei Verfahrensweisen: das Experiment und die Untersuchung:

> Während Experimente unter Laborbedingungen einer strengen Kontrolle im Hinblick auf die einflussnehmenden Variablen unterliegen, legen Untersuchungen den Schwerpunkt auf möglichst große Realitätsnähe in allen den Testaufbau betreffenden Faktoren, z.B. „echte“ Benutzer, realistische Größenverhältnisse bei der Testkollektion, „natürliche“ Formulierung der Aufgabenstellung.[50]

Bei der Planung eines Retrievaltests muss man sich mit verschiedenen praktischen Fragestellungen beschäftigen wie der Auswahl eines repräsentativen Test-Dokumentenbestands, der Formulierung der Test-Aufgaben, der Auswahl der Test-Personen oder auch der Koordination des gesamten Tests.[51] Es gibt aktuelle Evaluationsinitiativen wie TREC (Text REtrieval Conference), eine dynamische Plattform zur Evaluation von Retrievalverfahren, die das Ziel verfolgt, „umfangreiche, standardisierte Testkollektionen und Bewertungsprozeduren auf der

48 Vgl. Christa Womser-Hacker: „Theorie des Information Retrieval III: Evaluierung“, a.a.O., S. 229.
49 Ebd., S. 229-230.
50 Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 2: Glossar, a.a.O., S. 107.
51 Vgl. Christa Womser-Hacker: „Theorie des Information Retrieval III: Evaluierung“, a.a.O., S. 232.

Grundlage realistischer Anwendungsgebiete bereitzustellen".[52] Hier oder bei anderen Initiativen kann man sich für die Planung und Durchführung eines Retrievaltests Unterstützung holen.

Die Antworten auf die genannten praktischen Fragestellungen sind variabel. Um die Aussagefähigkeit von Retrievaltests bestimmen zu können, müssen diese Variablen für jeden Test definiert, kontrolliert und offen gelegt (dokumentiert) werden. Nohr listet die folgenden Gruppen von Variablen auf:

- Zweck des Tests
- technische Umgebung
- menschliche Umgebung
- Datensammlung
- Suchparameter
- Relevanz (Definition, Abstufung)
- Indexierung[53]

Durch diese Vielzahl von Einflussgrößen wird die Schwierigkeit der Vergleichbarkeit von Retrievaltests deutlich, auch wenn Initiativen wie die oben genannte TREC (oder GIRT für den deutschsprachigen Raum) einheitliche und kontrollierte Testbedingungen für vergleichende Retrievalexperimente zur Verfügung stellen.[54]

Nach der Theorie des Information Retrieval, die in die Grundlagen von Information Retrieval, Information-Retrieval-Systemen, automatischer Indexierung, Information-Retrieval-Modellen und Retrievaltests eingeführt hat, geht es in den folgenden Kapiteln um die Praxis des Information Retrieval im Allgemeinen sowie in ausgewählten Projekten.

52 Ebd., S. 232.

53 Vgl. Holger Nohr: *Grundlagen der automatischen Indexierung*, a.a.O., S. 157-158.

54 Zum geschichtlichen Abriss und untersuchten Fragestellungen von Retrievaltests sowie praktischen Hinweisen zur Durchführung vgl. Elisabeth Sachse, Martina Liebig und Winfried Gödert: *Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt*. Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft Bd. 14. Köln: Fachhochschule, Fachbereich Bibliotheks- und Informationswesen, 1998.

3 Praxis des Information Retrieval

Die Anwendung automatischer Indexierungs- und Retrievalverfahren kann in verschiedenen Kontexten erfolgen:

- organisationsinterne Verwaltung
- Medien- bzw. Pressedokumentation
- Internet
- Online-Datenbanken
- Bibliothek

3.1 Organisationsinterne Anwendung

Innerhalb von Unternehmen gewinnt Information Retrieval zunehmend an Bedeutung für die Verwaltung kleinerer, organisationsinterner Dokumentbestände. Diese enthalten in der Regel weniger Fachartikel, sondern setzen sich primär aus nicht-publizierten Dokumenten zusammen, die im Laufe der „Organisationsarbeit" entstehen, z.B. Briefe, Memos, Stellungnahmen, Gutachten, Sitzungsprotokolle oder Arbeitsberichte. Da solche Texte, die häufig bereits in elektronischer Form vorliegen, für die Organisation wichtiges Wissen enthalten, haben Methoden und Verfahren des Information Retrieval eine zentrale Bedeutung für organisationsinterne Informationssysteme. Reimer führt folgende Bedingungen für einen Einsatz im Bürobereich an:

- die Dokumente selbst sind im Volltext zu speichern (inkl. Multimedia-Fähigkeit)
- die Dokumente sind inhaltlich inhomogen
- die Diskursbereiche sind instabil, d.h. Erweiterungen um neue Begriffe, neue Bereiche oder neue Dokumenttypen kommen vor
- ein höherer Recall muss geleistet werden
- Benutzungsfreundlichkeit inkl. Modifikation von Profilen muss geboten werden[55]

Die betriebliche Informationswirtschaft geht aber noch weiter, indem über die Verwaltung hinaus auch die Bearbeitung von Dokumenten durch Information-Retrieval-Technologien unterstützt wird. Unternehmen wie Versicherungsanstalten oder Versandhäuser sortieren und bearbeiten ihre Dokumente anhand moderner Indexierungs- und Retrievalverfahren. Die eingehende Post wird digitalisiert (Scannen + OCR), sodass sie zusammen mit den eingegangenen Mails[56] nach den unterschiedlichen Inhalten bzw. Textsorten (Rechnungen, Reklamationen, Bestellungen etc.) sortiert und automatisch der jeweiligen Sachbearbeitung zugestellt werden kann. Hierbei gehören zu den technischen Maßnahmen die elektronische Archivierung, der Einsatz von Klassifikations- und Analysesoftware und ein komplexes Routing zu den Sachbearbeitern. Der Informations- und Dokumentenfluss innerhalb des Unternehmens soll automatisiert und gesteuert werden. Ziel ist die Einleitung eines „Workflow nach Posteingang".[57] Durch die genannten technischen Maßnahmen werden in den Unternehmen und Verwaltungen Verfahren des Information Retrieval relevant, die bislang auf die Literaturdokumentation mit ihren bibliographischen Fachdatenbanken bzw. den Volltextarchiven in der Pressedokumentation oder neuerdings auf die Informationsgewinnung aus dem Internet beschränkt waren.

---

55 Vgl. Ulrich Reimer: „Verfahren der automatischen Indexierung", a.a.O., S. 172-173.

56 Im Zusammenhang mit Mails kommen häufig Spamfilter und Antwortsysteme zum Einsatz, die ebenfalls automatische Indexierungs- und Retrievalverfahren verwenden.

57 Vgl. Holger Nohr: *Grundlagen der automatischen Indexierung,* a.a.O., S. 14.

3.2 Exkurs: Datenbank KURS zur Aus- und Weiterbildung

Bei der Datenbank KURS handelt es sich um eine Datenbank zur Aus- und Weiterbildung, die von der Bundesagentur für Arbeit herausgegeben und von einem privatrechtlichen Fachverlag redaktionell bearbeitet wird. Die Datenbank enthält Bildungsangebote aus Deutschland und dem angrenzenden Ausland aus allen Bildungsbereichen, z.B. Studiengänge, Umschulungen, Meisterausbildungen, IHK-Fortbildungen. Während der Projektlaufzeit von 1997 bis 2005, in der eine Mannheimer Projekt- und Verlagsgesellschaft mit der redaktionellen Bearbeitung betraut war, wurden die Angaben zu den Bildungsangeboten vom Verlag auf verschiedenen Wegen bei den Bildungseinrichtungen ermittelt. Diese Angaben beinhalteten Titel, Abschlussbezeichnung, Abschlussart, Inhalte, Kosten, Termine, Teilnahmevoraussetzungen uvm. Sie kamen auf verschiedenen Wegen und in unterschiedlichster Form im Verlag an, durchliefen aber alle denselben Workflow. Ziel war der papierlose Informationsfluss. Die redaktionelle Bearbeitung der Datenbank erfolgte mit einem web-basierten Redaktionssystem, das im Browser geöffnet wurde.

Angaben zu Bildungsangeboten in Printform (Broschüren, Flyer, Vorlesungsverzeichnisse etc.) wurden eingescannt und durchliefen anschließend die optische Zeichenerkennung sowie die Strukturerkennung (automatisierte Erkennung von Textstrukturen anhand des Layouts). Angaben zu Bildungsangeboten, die bereits elektronisch, aber unstrukturiert vorlagen (z.B. in Microsoft Word), durchliefen nur noch die Strukturerkennung. Daten, die bereits digital und strukturiert ankamen (z.B. Datenbanken von Kooperationspartnern oder Eingaben über ein Formular im Internet), konnten direkt in den Workflow eingespeist werden (so genannte Importdaten). Nach allen vorbereitenden Maßnahmen handelte es sich um XML-Files, die im weiteren Workflow bearbeitet wurden. Die folgende Grafik veranschaulicht die beschriebenen Wege:

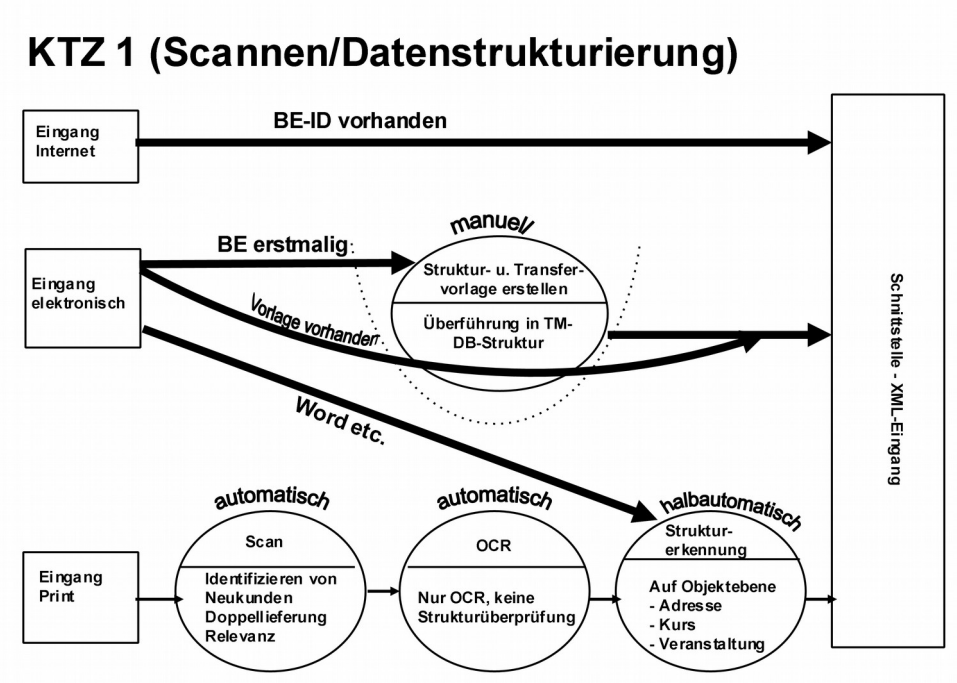

Abb. 3: Erzeugung der XML-Files im Kompetenzzentrum (KTZ) 1 nach Thomas Pfundstein[58]

58 Quelle: Schulungsunterlagen der ehemaligen Transmedia Projekt- und Verlagsges. mbH, Mannheim.

Prinzipiell gab es bei der Pflege der einzelnen Datenbankobjekte nur drei Möglichkeiten (Status): Neuanlagen, Löschungen und Änderungen, weshalb die Importdaten immer mit den bereits in der Datenbank vorhandenen Daten abgeglichen werden mussten, um den richtigen Status herauszufinden. Dieser Abgleich fand computerunterstützt statt, d.h. das Redaktionssystem schlug den Status vor (z.B. eine Zuweisung zwischen einem Importobjekt und einem Datenbankobjekt zwecks Änderung). Hierbei wurde eine halbautomatische Inhaltserschließung eingesetzt, die sowohl auf computerlinguistischen als auch auf statistischen Verfahren basierte. Der menschliche Datenbanksachbearbeiter/-redakteur konnte alle Vorschläge, die in einer Rangfolge (Ranking) nach der prozentualen Wahrscheinlichkeit ausgegeben wurden, manuell verwerfen und problematische Sachverhalte intellektuell klären, wie das folgende Schema zeigt:

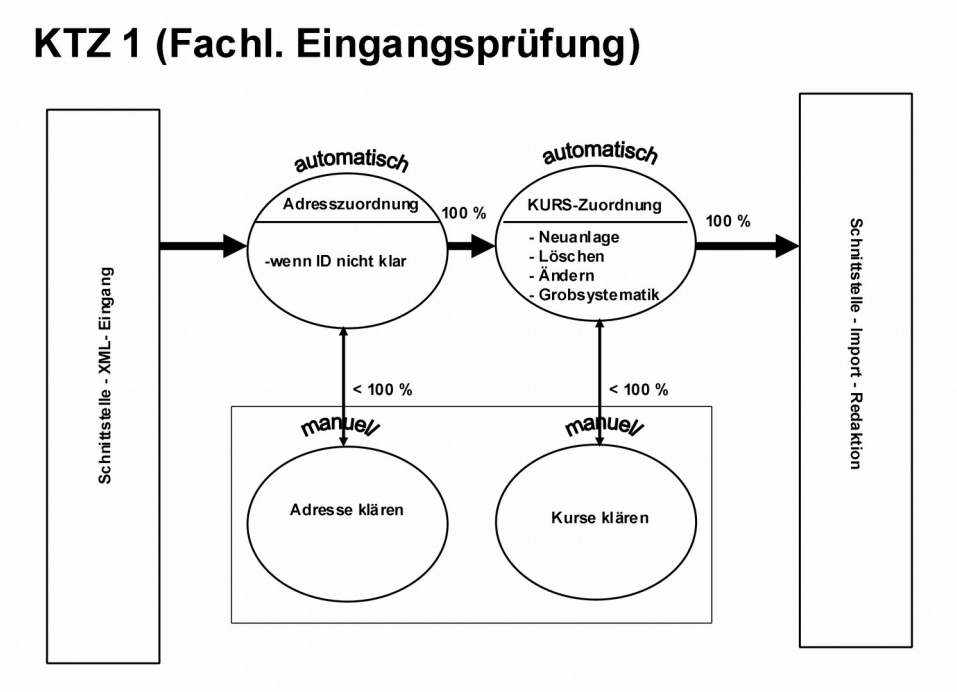

Abb. 4: Zuordnung von Importobjekten zu Datenbankobjekten nach Thomas Pfundstein[59]

Die Komponente der Inhaltserschließung, bei der es sich um verschiedene Softwarebausteine zur inhaltlichen Analyse von Textdokumenten einer Firma aus Saarbrücken handelte, kam an mehreren Stellen im Redaktionssystem zum Tragen: beim erwähnten Abgleich der Importobjekte (Adressen, Kurse) mit den Datenbankobjekten (Funktionalität der Ähnlichkeitssuche), bei der Klassifikation und bei der internen unscharfen Suche. Während die Ähnlichkeitssuche die Zuweisung von Objekten auf derselben Ebene unterstützte (z.B. importierte Adresse zu vorhandener Adresse), unterstützte die Klassifikation die Zuordnung zu übergeordneten Objekten, nämlich die Zuordnung von Bildungsangeboten zu einer Systematikposition. Die Systematik mit ihren einzelnen Positionen (Notationen oder Signaturen ähnlich) stellte das berufskundliche Ordnungssystem der Datenbankinhalte dar. Die Zuordnung zur übergeordneten Ebene erfolgte durch den Abgleich mit den sinntragenden Datenfeldern der untergeordneten Ebene. Im Gegensatz zur scharfen Suche (Datenbankabfragen über Boolesche Operatoren, die nur das Retrievalgewicht 0 oder 1 kennen) operierte die unscharfe Suche mit Ähnlichkeiten. Sie war dabei fehlertolerant, d.h. Tippfehler wurden zum Teil ignoriert (z.B. wurde bei der Eingabe von „Rostok“ auch nach „Rostock“ gesucht). An allen Stellen wurden die Treffer in einem Ranking ausgegeben; der menschliche Bearbeiter konnte einen der vorgeschlagenen

59 Quelle: Schulungsunterlagen der ehemaligen Transmedia Projekt- und Verlagsges. mbH, Mannheim.

Treffer übernehmen oder manuell etwas Eigenes zuweisen.

Durch den Einsatz moderner Indexierungs- und Retrievalverfahren bei der internen redaktionellen Bearbeitung der KURS-Datenbank sollten zum einen die subjektive menschliche Bearbeitung objektiviert bzw. vereinheitlicht und zum anderen Personalkosten eingespart werden.

3.3 Anwendung im Informations- und Dokumentationsbereich

Die Pressedokumentation mit ihrer täglich zu indexierenden Dokumentenmasse stellt den klassischen Anwendungsfall für automatische Indexierungsverfahren dar. Beispiele sind hier Gruner + Jahr mit dem Verfahren DocCat oder der Berliner Verlag mit GAdT (Grammatikalische Analyse deutschsprachiger Texte). Bei DocCat handelt es sich um ein System für die automatische Klassierung von Texten, bei dem eine Mischung aus statistischen, computerlinguistischen und begriffsorientierten Verfahren automatischer Indexierung angewandt wird. Pro Tag werden ca. 700 Artikel ohne technische Probleme bearbeitet. Der Einsatz dieses Verfahrens bringt eine Reduzierung von Routinetätigkeiten bei gleichzeitigem Ausbau von Bewertungs- und Korrekturarbeiten mit sich. GAdT, ein computerlinguistisches Verfahren, vergibt Terme aus einem in der Entwicklung befindlichen Thesaurus.[60]

Was den Kontext Internet betrifft, so verwenden Internet-Suchmaschinen (z.B. Google) sowie Bildsuchmaschinen Methoden des Information Retrieval. Suchmaschinen spielen eine zentrale Rolle in der Informationsrecherche. Ihr Erfolg gründet auf vier Faktoren: leichte Bedienbarkeit, Schnelligkeit der Suche, Umfang der erfassten Daten und Qualität des Rankings. Das Einsatzgebiet von Suchmaschinen beschränkte sich lange Zeit auf die Suche von unstrukturierten (Volltext-)Dokumenten aus dem World Wide Web. Mittlerweile stehen auch Lösungen für die Erschließung von strukturierten Datenbeständen zur Verfügung. Des Weiteren ist im Kontext Internet GERHARD (German Harvest Automated Retrieval and Directory) als ein Informationssystem zu nennen, bei dem begriffsorientierte Verfahren eingesetzt werden. Hier werden Internetquellen aus einem intellektuell erstellten Pool von Servern im Bereich der Fachinformation automatisch erschlossen. Dabei kommt eine Mischung aus Additions- und Extraktionsmethoden zum Tragen: Auf der Basis extrahierter Stichwörter werden den Internetadressen Notationen der Dezimalklassifikation zugeteilt.[61]

Information Retrieval findet ebenfalls bei der praktischen Informationssuche in qualitativ hochwertigen und gut strukturierten Online-Datenbanken statt, die über eine entsprechende Web-Oberfläche von den Nutzern erreicht werden können:

> Online-Datenbanken bieten qualitativ hochwertige Informationen zu allen Fach- und Wissensgebieten. Der Input wird von Fachleuten sorgfältig ausgewählt, redigiert und mit entsprechend vorgegebenen Regelwerken indexiert. Die so aufbereiteten Fachinformationen werden in strukturierten Datenbanken gespeichert, die eine effiziente Suche auch zu komplexen Fragestellungen erlauben. Die Web-Suchoberflächen erleichtern die Recherche und machen das Erlernen komplizierter Suchbefehle weitgehend überflüssig. Darüber hinaus bieten sie den Online-Zugriff zu Suchhilfen wie Datenbankbeschreibungen, Klassifikationen, Zeitschriftenlisten und Thesauri, die für eine professionelle Recherche zu komplexen Fragestellungen wichtig sind. Die Ausgabe der Rechercheergebnisse ist in den unterschiedlichen Formaten möglich, die graphische Aufbereitung wird durch entsprechende Werkzeuge unterstützt.[62]

60 Vgl. Jutta Bertram: *Einführung in die inhaltliche Erschließung*, a.a.O., S. 111.
61 Ebd., S. 110.
62 Joachim Kind: „Praxis des Information Retrieval“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch

3.4 Anwendung im Bibliotheksbereich

Anwendungsmöglichkeiten automatischer Indexierungs- und Retrievalverfahren ergeben sich im Kontext Bibliothek durch den Einsatz für die Literatursuche in Digitalen Bibliotheken und im OPAC.

Der Online-Katalog, in dem mittels eines Browsers zu jeder Zeit und an jedem Ort recherchiert werden kann, hat in den meisten wissenschaftlichen Bibliotheken den klassischen Zettelkatalog abgelöst. Da Zettelkataloge nicht mehr aktualisiert werden, bietet er dem Benutzer die einzige Möglichkeit, aktuelle Daten zu recherchieren. Er vereint die Funktionen des alphabetischen Katalogs mit dem Sachkatalog. Seine Verbreitung und Verfügbarkeit steht aber in Kontrast zu seiner Qualität. Grundlage der Daten ist nämlich nach wie vor die konventionelle bibliothekarische Titelaufnahme, die zumeist ergänzt wird durch die Schlagwörter der Schlagwortnormdatei. Ein OPAC ist immer durch den Kompromiss zwischen Endnutzeranforderungen und bibliothekarischer Indexierung geprägt. Der vorhandene Datenbestand ist nicht durchgängig verbal erschlossen; eine retrospektive intellektuelle Erschließung ist aus Zeit- und Kostengründen nicht durchführbar. Die Zahl der nach RSWK vergebenen Deskriptoren ist vergleichsweise gering. Die RSWK als langjähriges Instrument klassischer bibliothekarischer Sacherschließung sollten mit der Neuauflage von 1998 dem Einsatz in Online-Katalogen angepasst werden. Letztlich wurde mit dem Festhalten an dem Regelwerk auch für den Online-Katalog diesem Instrument gegenüber der Einführung maschineller Verfahren der Vorzug gegeben. Auch die Titelbeschreibung selbst ist als Datengrundlage nach Lohmann problematisch: Titeldaten geben aufgrund ihres geringen Umfangs in vielen Fällen nicht in ausreichendem Maße Auskunft über den Inhalt des zugrunde liegenden Werks. „Sowohl für die Suche als auch für den Entscheidungsprozeß bei der Titelauswahl ist die Titelbeschreibung aufgrund ihrer Informationsarmut insofern ein unzureichendes Mittel.“[63] Während Bibliotheken lange Zeit das Konzept der konventionellen Titelaufnahme verfolgten, orientierten sich moderne Informationssysteme zunehmend am Volltext, zumindest aber an reichhaltig erschlossenen Dokumenten.

Auch die Retrievalumgebung des typischen Bibliotheks-OPAC entspricht nicht den technisch möglichen Funktionalitäten. Der Nutzer sucht hier nach wie vor mit Booleschen Operatoren; die Ausgabe der Treffer erfolgt nicht nach Relevanz, sondern nach dem Erscheinungsjahr, sodass nicht zwangsläufig die besten Treffer an erster Stelle stehen. Da sich die meisten Nutzer nur die ersten Treffer der Ergebnisliste anschauen, bleiben relevante Treffer eventuell unbemerkt. Aber auch die im OPAC vorhandenen Funktionalitäten werden von den Nutzern aus Unkenntnis nicht ausgeschöpft. Ein großer Teil der Endnutzer ist weder in der Lage, eine Autor-Stichwort-Verknüpfung kompetent auszuführen noch vorhandene weiterführende Rechercheinstrumente (z.B. Einschränkung nach Materialart oder Erscheinungsjahr) einzusetzen. Die Sachrecherche des Endnutzers erfolgt deshalb fast ausschließlich über die „Quick-and-Dirty-Methode“ der Stichwortrecherche.[64] Die Ergebnisse solcher Suchanfragen sind oft unbrauchbar, weil sie gekennzeichnet sind durch übergroße Treffermengen, leere Treffermengen und sehr kleine Treffermengen. Übergroße Treffermengen sind das Ergebnis von Suchanfragen mit zu wenig spezifischen Suchbegriffen (z.B. Deutschland oder Geschichte). Der Nutzer kann die Treffer nicht mehr überschauen; manchmal übersteigt die Größe der Ergebnisliste sogar die Kapazität des Rechners, sodass z.B. nur die ersten 300 Treffer angezeigt werden.

(Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, a.a.O., S. 398.

63 Hartmut Lohmann: *KASCADE: Dokumentanreicherung und automatische Inhaltserschließung,* a.a.O., S. 11.

64 Ingrid Recker, Marc Ronthaler, Hartmut Zillmann: „OSIRIS – ein Hyperbase Front End System für OPACs“. In: Bibliotheksdienst 30. Jg. (1996), H. 5. S. 833.

Leere Treffermengen entstehen, wenn der Suchterminus nicht mit dem Indexierungsterminus übereinstimmt, was an Eingabefehlern (z.B. Tippfehlern) oder an der speziellen Ansetzung von Schlagwörtern liegt. Eine Fehlerkorrektur bei der Eingabe findet in der Regel nicht statt, ebenso wie eine natürlichsprachige Eingabe zumeist nicht möglich ist. Kleine Treffermengen scheinen auf den ersten Blick perfekt zu sein, weil es zum einen überhaupt Treffer gibt und zum andern die Menge handhabbar ist, aber sie sind am gefährlichsten, weil sie den Nutzer täuschen können. Während übergroße oder leere Treffermengen den Nutzer darauf hinweisen, dass er seine Anfrage modifizieren muss, lassen ihn kleine Treffermengen im Glauben, dass seine Suchanfrage ideal formuliert war. Es kann aber sein, dass sein Ergebnis überhaupt nicht vollständig ist, dass er also längst nicht alle für seine Sachrecherche eigentlich relevanten Literaturhinweise mit seiner Suchanfrage erhalten hat. Der Nutzer erreicht weder die Titel, in denen das gesuchte Thema in einem Kapitel abgehandelt wird, noch das Titelmaterial anderer Sprachen. Problematisch sind z.B. auch Komposita, unregelmäßige Pluralbildungen, flektierte Formen oder Synonyme. Die meisten Nutzer kennen sich zu wenig mit Trunkierungs- bzw. Maskierungsmöglichkeiten aus. Durch die Anreicherung des Titelmaterials mit zusätzlichen Begriffen z.B. aus Inhaltsverzeichnissen besteht die Möglichkeit, das am meisten benutzte Stichwort-Retrieval zu verbessern.

Die geschilderte Situation zeigt, dass noch ein erheblicher Spielraum für die nutzerorientierte Verbesserung der Literatursuche über den OPAC besteht. Die Verbesserung der Retrievalsituation wird durch moderne Verfahren der Kataloganreicherung (sog. Catalogue Enrichment) mittels Scannen von Titelblättern, Klappentexten, Inhaltsverzeichnissen und Registern, automatischer Indexierung der Datengrundlage sowie maschineller Retrievalverfahren (Fehlerkorrektur, Relevance Ranking, Relevance Feedback etc.) in verschiedenen Projekten angestrebt. Nach Bertram lassen sich im Rahmen der automatischen Indexierung besonders computerlinguistische Verfahren im Bibliotheksbereich wirksam einsetzen:

> Daß im Bibliotheksbereich gerade computerlinguistische Verfahren ihr Potential sehr gut entfalten können, liegt vermutlich in der Art der zugrundeliegenden Textmenge begründet. Denn da diese selbst bereits zu einem wesentlichen Teil auf Inhaltsverdichtungen beruht (z.B. durch die Berücksichtigung von Titeln und Inhaltsverzeichnissen), sind Termgewichtungen hier nicht so vonnöten wie etwa im Pressebereich. Die gescannten Terme sind per se zumeist 'gute' (entscheidungsstarke, relevante) Terme.[65]

Einige Projekte aus dem deutschsprachigen Raum aus jüngerer Vergangenheit und Gegenwart, in denen mit dem Einsatz maschineller Verfahren auf der Erschließungs- wie der Retrievalebene neue Wege außerhalb der klassischen intellektuellen Indexierung beschritten wurden, werden im Folgenden vorgestellt.

3.4.1 Rückblick

3.4.1.1 OSIRIS

OSIRIS (Osnabrück Intelligent Research Information System) wurde gemeinsam von der Universitätsbibliothek Osnabrück und dem Institut für Semantische Informationsverarbeitung an der Universität Osnabrück entwickelt. Die Deutsche Forschungsgemeinschaft (DFG) förderte seit 1996 im Rahmen eines dreijährigen Projektes die Entwicklung von Teilkomponenten.

65 Jutta Bertram: *Einführung in die inhaltliche Erschließung*, a.a.O., S. 110.

> OSIRIS ist ein multilinguales intuitiv-natürlichsprachlich zu benutzendes Retrievalsystem insbesondere für bibliographische Datenbanken, das die Anwendung klassischer Recherchetechniken (spezielle Kommandos, Boolesche Verknüpfungen, Trunkierung etc.) für den Benutzer überflüssig macht und im Vergleich zu einem konventionellen OPAC (Online Public Access Catalog) qualitativ bessere Suchergebnisse erzielt.[66]

Das OSIRIS-Projekt hatte als Zielsetzung, auf der Basis des vorhandenen Daten- und Informationsbestands in Form eines „Intelligenten User Interface" deutliche qualitative Verbesserungen im Hinblick auf Formal- und Sachrecherchen des Endnutzers zu erbringen. Durch eine englischsprachige Oberfläche, die die Sachrecherche mit englischsprachigen Suchbegriffen umfasst, sollte der OPAC „Internationalität" erhalten. Zugleich entstand als Nebenprodukt eine Komponente zur Unterstützung der Fachreferatsarbeit, das so genannte Computer Aided Indexing (CAI).[67] OSIRIS sieht einen dreifach gestuften Endnutzerzugang zur OPAC-Datenbank vor:

- Formalrecherche
- Sachrecherche
- Expertenmodus

Die Formalrecherche deckt die gängigen Anfragen im Bereich der Titelsuche ab. Bei der Analyse der Titelstichwörter kommt bereits eine Morphologiekomponente zum Einsatz. An der Eingabeschnittstelle zur Sachrecherche ist auf der Eingabemaske ein vorgegebener Satz zu vervollständigen („Ich suche Literatur zum Thema..."). Es ist erforderlich, die Ergänzung des Benutzers syntaktisch und semantisch zu analysieren, also zu parsen und inhaltlich auf die Terminologie der Klassifikation und weiterer begrifflicher Anreicherungen abzubilden. Es wird ein kleiner, auf Teilsätze ausgelegter, robuster und effizienter Parser benötigt, der die wesentlichen Fälle bearbeiten kann, aber auch einzelne Fälle zur Präzisierung zurückgeben darf. Die eingegebenen Teilsätze (meistens Nominalphrasen) müssen vom System verstanden werden. Bestimmte Eingabeteile wie Zeit-, Sprach- oder geographische Aspekte werden anhand von Schlüsseltabellen erkannt. Das OSIRIS-Basisvokabular wird im Wesentlichen aus der OPAC-Datenbank selbst gewonnen. Aber auch aus den Klassenbezeichnungen der Systematik werden Suchbegriffe abgeleitet. Der deutsche Wortindex z.B. wird aus den aus der Systematik abgeleiteten Suchbegriffen und aus den extrahierten RSWK-Schlagwörtern aufgebaut. Weitere Komponenten sind Browsing in der Systematik, selbstlernende Sachrecherche, Robustheit gegenüber Schreibfehlern, konstruktive Rückfragen (mit Vorschlägen) und statistische Bewertung der Suchergebnisse (durch Signifikanzzähler in den OSIRIS-Tabellen).

Die Vorteile von OSIRIS sind zum einen die robuste, natürlichsprachige Benutzerschnittstelle, zum anderen die intelligente, automatische Aufbereitung des verfügbaren Datenbestands in einer Wissensbasis. Unter den computerlinguistischen Bausteinen sind der Parser und das Lexikon von besonderer Bedeutung. Um die Effizienz von OSIRIS abzuschätzen, wurde ein Retrievaltest durchgeführt, bei dem die Retrievalergebnisse unter OSIRIS mit denen des OPAC der UB Osnabrück verglichen wurden:

> Im Durchschnitt hat die OSIRIS-Retrievalkomponente im Vergleich zum OPAC der Universitätsbibliothek Osnabrück einen um den Faktor 11 größeren Recall. Wie das jeweilige Trefferbild im OSIRIS-System zeigt, sind die Suchergebnisse außerdem von hoher Präzision.[68]

---

66 „OSIRIS – Osnabrück Intelligent Research Information System". In: ABI-Technik 20, 2000, Nr. 1. S. 89.

67 Vgl. Ingrid Recker, Marc Ronthaler, Hartmut Zillmann: „OSIRIS – ein Hyperbase Front End System für OPACs", a.a.O., S. 834.

68 Marc Ronthaler, Hartmut Zillmann: „Literaturrecherche mit OSIRIS. Ein Test der OSIRIS-Retrievalkomponente". In: Bibliotheksdienst 32. Jg. (1998), H. 7. S. 1208.

3.4.1.2 MILOS I und II

Die DFG-geförderten Projekte MILOS I und II (Maschinelle Indexierung zur erweiterten Literaturerschließung in Online-Systemen) wurden in den Neunziger Jahren an der Universitäts- und Landesbibliothek (ULB) Düsseldorf durchgeführt. In ihnen wurde die automatische Indexierung erstmalig auf rein bibliothekarische Titeldaten in einem inhaltlich stark heterogenen Bestand angewandt. Das eingesetzte Indexierungssystem IDX wurde von Prof. Dr. Harald H. Zimmermann, Fachrichtung Informationswissenschaft der Universität des Saarlandes, entwickelt. Es handelt sich um ein rein wörterbuchbasiertes Verfahren, d.h. alle Arbeiten am Text beruhen auf einem Abgleich mit verschiedenen elektronischen Wörterbüchern. Dabei unterstützt IDX folgende Funktionen für die Sprachen Deutsch, Englisch und Französisch:

- Ermittlung von Grundformen zu den im Text vorkommenden Wortformen (Bibliotheken -> Bibliothek)
- Markierung bzw. Eliminierung von Stoppwörtern
- Bereitstellung von Wortableitungen und von (sinnvollen) Bestandteilen von Komposita (bibliothekarisch -> Bibliothek; Bibliothekswissenschaft -> Bibliothek, Wissenschaft)
- Bereitstellung von Begriffsrelationen (Stichwort -> Schlagwort; Begriff -> Oberbegriff; Begriff -> verwandter Begriff)
- Mehrwort-Erkennung und Wortbindestrichergänzung (Regeln für den Schlagwortkatalog; Buch- und Bibliothekswesen -> Buchwesen, Bibliothekswesen)

Seit Januar 1994 wurde mit MILOS I (1993/94) und II (1995/96) untersucht, welche Möglichkeiten für den Einsatz von IDX in Bibliotheken bestehen. Ziele von MILOS I waren die Weiterentwicklung von IDX und die Anpassung des Systems an die spezielle Arbeitsumgebung einer wissenschaftlichen Universalbibliothek. Während der einjährigen Laufzeit des Projekts wurden in großem Umfang Titeldaten der ULB Düsseldorf automatisch indexiert, aus dem praktischen Einsatz heraus zahlreiche Systemverbesserungen programmiert und auf der Grundlage der indexierten Daten neue Wörterbücher aufgebaut bzw. bereits bestehende stark erweitert. Ein abschließender Retrievaltest mit automatisch erzeugten Indexaten führte zu durchweg positiven Ergebnissen, so dass die ULB die automatische Indexierung als festen Bestandteil der Suchmöglichkeiten in ihren OPAC integriert hat. Die Ergebnisse von MILOS I haben zu einem Produktpaket MILOS geführt, das die Funktionalität der automatischen Indexierung für die drei Sprachen Deutsch, Englisch und Französisch für Nachnutzer verfügbar macht. Aufbauend auf den Ergebnissen von MILOS I wurde innerhalb des Folgeprojektes MILOS II seit Januar 1995 an einer Ausweitung der Funktionalität von IDX gearbeitet. Ziel von MILOS II war die sinnvolle Zusammenführung von konventionellen Methoden der inhaltlichen Erschließung - verbale Sacherschließung nach RSWK, klassifikatorische Erschließung - mit den Möglichkeiten eines automatischen Indexierungsverfahrens. In Zusammenarbeit mit der Deutschen Nationalbibliothek (zu dem Zeitpunkt noch deutsche Bibliothek) wurde IDX durch die Einbindung von Thesaurusrelationen der Schlagwortnormdatei in das Wörterbuchkonzept um semantische Funktionalitäten erweitert (elektronischer Thesaurus). Die für die automatische Indexierung im neuen Funktionsumfang notwendige Erweiterung der Wörterbücher erfolgte im Sinne größtmöglicher Nachnutzbarkeit auf der Grundlage der maschinenlesbaren Titeldaten der Deutschen Nationalbibliothek.[69]

69 Die Kurzbeschreibung der Projekte MILOS I und II wurde der Homepage der ULB Düsseldorf entnommen. http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/milos/mil_kurz [Letzter Aufruf: 01.05.2007]

Nach dem erfolgreichen Abschluss von MILOS II und der zwischenzeitlich erfolgten Bewährung des MILOS-Systems im Routinebetrieb wurde die Praxis der inhaltlichen Erschließung an der Universitäts- und Landesbibliothek seit Beginn des Jahres 1997 auf ein Mischverfahren mit intellektueller und automatischer Komponente umgestellt. Ziel verbaler Erschließung sollte die Ergänzung des Titelvokabulars um indexierungstaugliche Begriffe sein. Deutschsprachige Literatur sollte grundsätzlich durch automatische Indexierung bearbeitet werden; für fremdsprachige Literatur war die Vergabe von freien deutschsprachigen Deskriptoren vorgesehen. Diese wurden arbeitsteilig vom Fachreferat und von der Abteilung Inhaltserschließung gewonnen, wobei bereits verfügbare Stichwörter (ggf. in Übersetzung von z.B. Titelstichwörtern bzw. vorhandenem fremdsprachigen Erschließungsvokabular) hinzugezogen werden sollten. Eine Verschlagwortung nach den RSWK war nicht mehr vorgesehen; die Mitarbeit an der Pflege der Schlagwortnormdatei war jedoch auch weiterhin wichtiger Bestandteil der Inhaltserschließung.[70]

3.4.1.3 KASCADE

Ziel von KASCADE (Katalogerweiterung durch Scanning und Automatische Dokumenterschließung), dem Nachfolgeprojekt zu MILOS I und II, war die Erweiterung konventioneller bibliothekarischer Titeldaten um zusätzliche inhaltsrelevante Informationen und deren automatische Erschließung. Die so erreichte umfassende verbale und klassifikatorische Erschließung der Dokumente schuf die Basis für den Einsatz fortgeschrittener Navigations- und Suchverfahren auf der Retrievalseite. Zur Umsetzung des Vorhabens fand eine Zusammenarbeit mit der Juristischen Fakultät der Heinrich-Heine-Universität Düsseldorf sowie mit der Fachrichtung Informationswissenschaft an der Universität des Saarlandes, Saarbrücken, statt. Es wurden - exemplarisch für das Fachgebiet Jura - Titelaufnahmen durch den Einsatz von Scanningverfahren um die Inhalte von Inhaltsverzeichnissen angereichert. Darüber hinaus wurden verfügbare elektronische Volltexte in die Datenbasis übernommen. Erweiterte Titelaufnahmen und elektronische Volltexte wurden automatisch indexiert und im Rahmen einer erweiterten maschinellen Indexierung automatisch klassifiziert (sog. Themen-Aspekt-Identifizierung (THEAS) für die Erkennung von Themen-Aspekt-Beziehungen in Mehrwortgruppen). Bei der automatischen Indexierung wurden die computerlinguistischen Verfahren aus MILOS um statistische Verfahren ergänzt, nämlich durch eine selektive automatische Indexierung zur gewichteten Extraktion von Deskriptoren (SELIX).

Die Anreicherung der Titeldaten umfasste Scanning von Inhaltsverzeichnissen von ca. 3.000 Titeln aus dem Bestand Jura, OCR, MILOS-Rechtschreibkontrolle und die Datenablage in MS-Access. Die automatische Indexierung beinhaltete die folgenden Schritte:

- MILOS-Indexierung mit Grundformermittlung und Dekomposition für Sachtitel, Schlagwörter und Volltexte der Inhaltsverzeichnisse
- SELIX-Gewichtungsindexierung in insgesamt fünf Programmschritten
- zusätzliche MILOS II-Indexierung
- Verknüpfen von Titelaufnahmen, Scanningresultaten und Indexierungsergebnissen und Ablage in Datenbank
- Generierung von THEAS-Relationen
- Aufbau der KASCADE-Testdatenbank

70 Vgl. Abschlussbericht zum Projekt MILOS II: http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/milos/mil_ber [Letzter Aufruf: 01.05.2007]

Der im Projekt erstellte Datenbestand wurde über ein konzeptorientiertes Information-Retrieval-System verfügbar gemacht. Auf dieser Basis wurde im Rahmen der Projektevaluation ein Retrievaltest durchgeführt. Positiv waren sehr gute Recall-Werte bei gleichzeitig sehr guten Precision-Werten. Negativ war, dass die Anreicherung von Titelaufnahmen (bestehender Bestände) durch Scanning in technischer Hinsicht und hinsichtlich des Arbeitsablaufs hochgradig problematisch war, und dass die für KASCADE/SELIX ermittelten Werte im Retrievaltest unter den Erwartungen lagen. Fraglich blieb, ob die Anreicherung von Titelaufnahmen zu einer Verbesserung der Suchergebnisse führte und ob das Retrieval auf angereicherte Titelaufnahmen durch die selektive automatische Indexierung verbessert wurde.[71]

3.4.2 Aktuelle Projekte

Aktuelle Projekte beschäftigen sich mit unterschiedlichen Formen der Anreicherung von Katalogdaten. Hierzu gehören beispielsweise Recommendersysteme, die aus Online-Shops (z.B. Online-Buchhandlungen wie www.amazon.de) bekannt sind. Recommendersysteme, die sich in verhaltensbasierte und explizite Dienste einteilen lassen, geben Empfehlungen. Bei den verhaltensbasierten Recommenderdiensten handelt es sich um automatische, aus statistischen Daten generierte Empfehlungen im Sinne von „Kunden, die diesen Titel aufgerufen bzw. ausgeliehen haben, haben auch folgende Titel aufgerufen bzw. ausgeliehen“ (z.B. im OPAC der Universitätsbibliothek Karlsruhe). Zu den expliziten Recommenderdiensten zählen Ranking- und Reviewdienste, die von Personen verfasst werden, wie Online-Rezensionen und Kundenbewertungen (z.B. im OPAC der Universitätsbibliothek Mannheim). Solche Dienste unterstützen nicht nur die Nutzer bei der Literaturrecherche, sondern können auch den Bestandsaufbau und die Sacherschließung verbessern. Sie stellen eine spezielle Form der bibliographischen Anreicherung dar.

Weiterhin aktuell sind Projekte, in denen zusätzliche Informationsquellen wie Inhaltsverzeichnisse eingescannt, spracherkannt sowie maschinell indexiert werden. Hier ist zum einen das Suchportal für wissenschaftliche Literatur dandelon.com, das von verschiedensten Einrichtungen, darunter auch Hochschul- und Landesbibliotheken, gemeinsam getragen wird, zu nennen. Auch bei dandelon.com werden mittels Scanning, OCR und maschineller Indexierung die Inhaltsverzeichnisse von Büchern (u.a.) erschlossen. Zum anderen gibt es das 180-T-Projekt vom Hochschulbibliothekszentrum (HBZ) des Landes Nordrhein-Westfalen in Köln, in dessen Verlauf die Inhaltsverzeichnisse von ca. 180.000 Büchern eingescannt wurden. Diese beiden Projekte bzw. Ansätze werden in den folgenden zwei Kapiteln dargestellt.

71 Vgl.Klaus Lepsky: „KASCADE“: http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/kascade/kas_proj [Letzter Aufruf: 01.05.2007]

4 Dandelon.com

Dandelon.com ist ein Suchportal für wissenschaftliche Literatur auf Basis der multilingualen, semantischen Suchmaschine intelligentSEARCH, IC INDEX und intelligentCAPTURE. Es bietet maschinell erschlossene Inhaltsverzeichnisse, Klappentexte, Zusammenfassungen aus Büchern, Artikel als Referenzinformation oder mit dem Volltext, sofern Zugangsrechte bestehen, sowie Websites wissenschaftlicher Forschungsinstitute und sonstiger interessanter Seiten. Dahinter steht ein internationales Netzwerk von Bibliotheken, Bibliotheksservicezentren, Verlagen, Buchhandel, Fachinformations- und Dokumentationszentren, Softwareentwicklung und Hochschulen. Sie alle erfassen, erschließen, teilen und suchen Informationen und Wissen über dandelon.com.

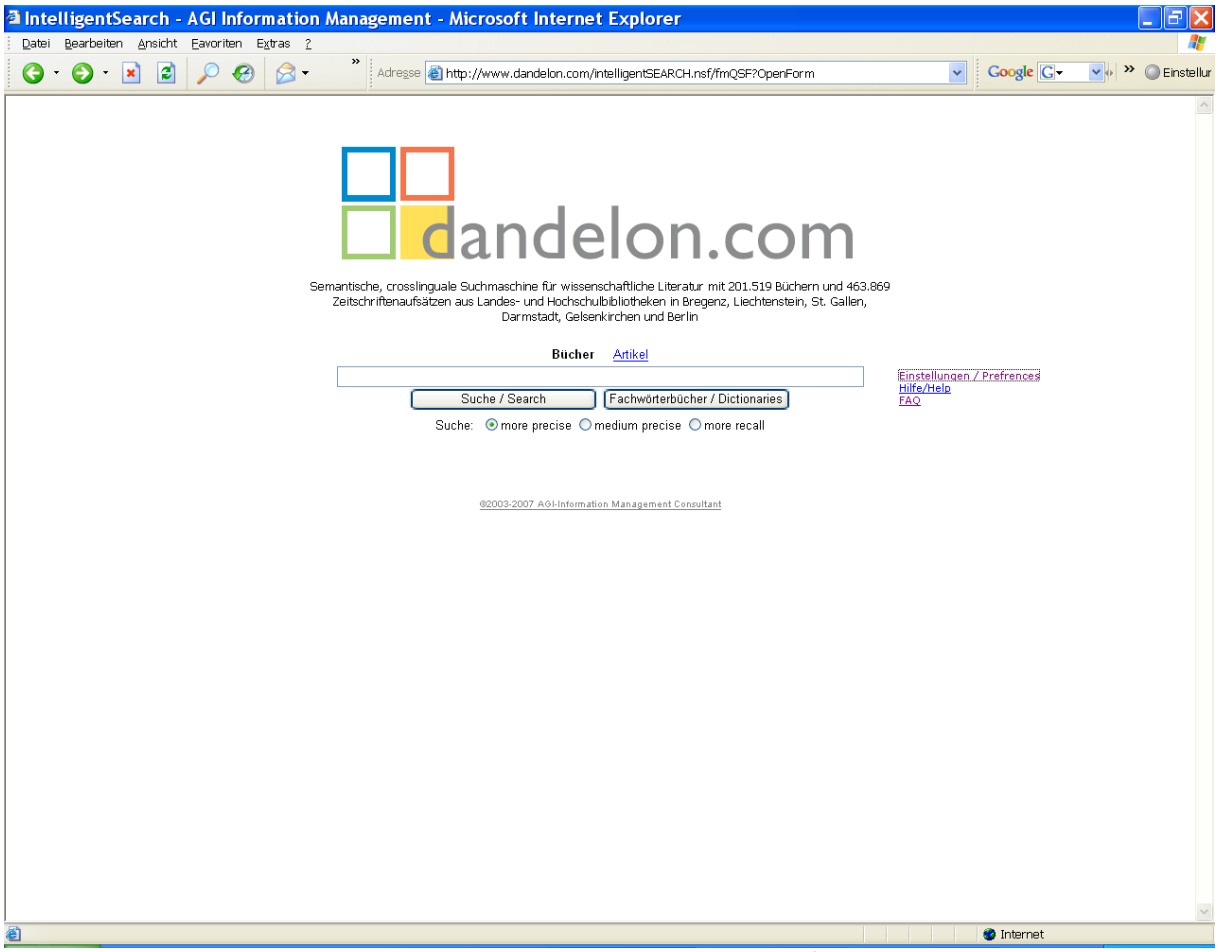

Abb. 5: Startseite von dandelon.com[72]

4.1 Projektziel

Das Ziel von dandelon.com ist die Verbesserung der thematischen Recherche durch die Zusammenführung von angereicherten Bibliothekskatalogen, Aufsätzen, Websites und Verlagsdaten. Das Ergebnis einer Sachrecherche soll exakter und vollständiger ausfallen als es bisher bei der Suche in OPACs ohne weitere Kataloganreicherung war. Durch ein einfaches Layout, das an Internetsuchmaschinen denken lässt, soll den Gewohnheiten der Nutzer entsprochen werden.

72 http://www.dandelon.com/intelligentSEARCH.nsf/fmQSF?OpenForm [Letzter Aufruf: 01.05.2007]

4.2 Projektpartner

Das Suchportal wird von National-, Landes- und Hochschulbibliotheken, dem Bibliotheksservice-Zentrum GBV, nationalen und internationalen Dokumentationszentren, Verlagen (Springer u.a.), der Zeitschriftenagentur SWETS, dem Buchhandel (Missing Link) und - als Initiator AGI-Information Management Consultants - gemeinsam getragen. Bisher wirken Partner aus vier europäischen Staaten in diesem Gemeinschaftsprojekt mit. Im Einzelnen sind dies: die Vorarlberger Landesbibliothek in Bregenz, die FH Burgenland, die Liechtensteinische Landesbibliothek, die Universität St. Gallen, die TU Darmstadt, die FH Gelsenkirchen, die FHTW Berlin, die TIB/UB Hannover und die SUB Hamburg. Missing Link ist eine in Bremen ansässige Versandbuchhandlung, die gewählt wurde, weil es sich bei ihr um einen Importeur ausländischer Titel für viele große Bibliotheken und sehr viele Buchhändler handelt, der auf wissenschaftliche Literatur fokussiert ist und einen großen Umfang lieferbarer Titel hat.[73] Bei AGI-Information Management Consultants (AGI-IMC) handelt es sich um das Unternehmen von Dipl.-Informationswissenschaftler Manfred Hauer in Neustadt/Weinstraße. Weitere Suchlösungen von AGI-IMC sind das Portal Informationswissenschaft, das Landtagsinformationssystem sowie das Forschungsinformationssystem des Landes Vorarlberg.

4.3 Projektorganisation

AGI-IMC als Dienstleister für die Erfassung der Daten stellt den teilnehmenden Bibliotheken die Software intelligentCAPTURE und das Personal zur Verfügung. Bei Bedarf können über AGI-IMC auch die kompletten mobilen Scanstationen bezogen werden. Die Durchführung findet vor Ort in den Bibliotheken statt, entweder im Dauerbetrieb integriert wie an der TU Darmstadt oder in Projektform wie an der TIB/UB Hannover. In Hannover wurden z.B. von November 2006 bis Januar 2007 durch studentische Hilfskräfte 20.000 Konferenzbände an drei Scanstationen mit der intelligentCAPTURE-Software erfasst, um den Katalog anzureichern. Die Daten werden im GBV-Katalog, im TIB-Katalog und bei dandelon.com sichtbar sein. Für dandelon.com sollen die PDF-Dateien noch einmal komplett neu berechnet werden, um sie kleiner und besser lesbar zu machen (keine Images im Vordergrund). Dieser Schritt erfolgt im Frühjahr/Sommer 2007. Mitte Februar fand die offizielle Übergabe der Daten an die TIB statt. Das Projekt wurde vom GBV in Göttingen im Rahmen der Initiative zur Erfassung von Inhaltsverzeichnissen finanziert. Im Rahmen dieser Initiative ist im Februar 2007 an der SUB Hamburg ein Projekt zur Erfassung von 42.000 Büchern in 32 Sprachen gestartet worden. Thematisch handelt es sich um die Sondersammelgebiete Portugal und Spanien und um sozialwissenschaftliche Literatur. Hier kommt erstmals die mobile Scanstation direkt zwischen den Regalen zum Einsatz. Sie steht über WLAN im Datenaustausch mit dem PICA-Bibliothekssystem und dem Server von dandelon.com.[74] Am 27. März 2007 startete ein Catalogue-Enrichment-Projekt an der UB Braunschweig, bei dem 10.000 häufig genutzte Titel von mindestens 150 Seiten Umfang vorwiegend aus den Sozial- und Technikwissenschaften bearbeitet werden.

---

73 http://www.missing-link.de [Letzter Aufruf: 01.05.2007]

74 Vgl. Mitteilung von Manfred Hauer: http://www.agi-imc.de/icapture/FAQ_in_iS.nsf/ec4bd625eb29e0cdc1256ce600385e4b/a07cf70e67256045c125718d0060f8dd?OpenDocument [Letzter Aufruf: 01.05.2007]

Der unbefristete Kooperationsvertrag zwischen AGI-IMC und dem GBV, der am 17.03.2005 auf dem Deutschen Bibliothekartag in Düsseldorf unterzeichnet wurde, stellt langfristig das Hosting von dandelon.com und die Links auf PDF-Dateien aus Inhaltsverzeichnissen im GBV-Katalog sicher.

Wenn über dandelon.com ein Titel gefunden wurde, der nicht ausleihbar ist, gibt es einen Link zur Bestellung bei der Bremer Buchhandlung. Aus intelligentCAPTURE heraus können Bibliotheken über eine integrierte Dokumentenlieferfunktion (Scan/Send) auch Dokumente direkt versenden.Verlage (z.B. Springer) und Agenturen liefern Daten, die in dandelon.com importiert werden.

4.4 Projektverlauf

Die Basis-Software von dandelon.com wurde bereits in den Neunziger Jahren im Rahmen von elektronischen Nachrichten-Filtering- und Verteilsystemen für Pressearchive eingesetzt, z.B. bei GENIOS oder Henkel. Die Bausteine hießen hier IC News als ein Produktions-, Archivierungs- und Verteilsystem für Unternehmen bzw. IC Individual News für die individualisierte Nachrichtennutzung. IC News (IC steht für Information Center) wurde seit 1994 bei einem Dutzend Kunden eingesetzt; der Baustein IC Individual News kam 1997 dazu. Ab 1999 bot AGI-IMC das Produkt IC News für Verlagshäuser an, mit dem gleich nach dem Satz am Abend die Zeitung komplett „zerschnitten“ und elektronisch bereitgestellt werden konnte, was die Erstellung eines Pressespiegels in den Unternehmen vereinfachte. Daraus entstand das Nachrichtenportal intelligentNEWS, an dem Factiva, Dialog NewsEdge, Presse-Monitor-Gesellschaft und Swets Blackwell beteiligt waren.[75]

Was den Einsatz der Software in Bibliotheken betrifft, so führte die Vorarlberger Landesbibliothek in Bregenz zusammen mit AGI-IMC im Jahr 2002 ein Projekt der Kataloganreicherung durch. Von März bis November wurden 4.000 Neuerwerbungen an einem Arbeitsplatz verarbeitet und geprüft. Durch zwei weitere Arbeitsplätze wurde im Jahr 2003 die Leistung gesteigert, sodass 12.000 Neuerwerbungen erfasst werden konnten. Hinzu kamen Volltexterschließungen aus dem Sondersammelgebiet Vorarlberg. Aus diesem Projekt entstand zunächst das Produkt intelligentCAPTURE, dann 2003 das Produkt intelligentSEARCH und im Frühjahr 2004 der öffentliche Service dandelon.com.

Ende Januar 2007 erreichte das Suchportal "dandelon.com" die Menge von 180.000 Buchtiteln, deren Inhaltsverzeichnisse überwiegend gescannt und maschinell indexiert wurden. Die Suche wurde um 1,5 Mio. Fachbegriffe in bis zu 20 Sprachen automatisch semantisch erweitert und übersetzt. Neu war die direkte Anzeige der Suchworte als PDF.[76] Mitte März 2007 wurden 200.000 Titel im Suchportal erreicht, woran die aktuellen Catalogue-Enrichment-Projekte der SUB Hamburg und der TIB Hannover großen Anteil hatten. Täglich kommen rund 1.000 Inhaltsverzeichnisse dazu.[77] Aktuell wird an der Vertonung von Volltexten geforscht, um Texte hören statt lesen zu können.

---

75 Vgl. Manfred Hauer: „Strukturierung, Erschließung und Präsentation von Nachrichtentexten“. In: Wissensmanagement: Strategie, Prozesse, Communities (Tagungsband 2002), S. 101-108. http://www.agi-imc.de/internet.nsf/RahmenDeutsch?OpenFrameSet [Letzter Aufruf: 01.05.2007]

76 Vgl. Mail von Manfred Hauer an inetbib@ub.uni-dortund.de vom 27.01.2007 http://www.ub.uni-dortmund.de/listen/inetbib/msg32251.html [Letzter Aufruf: 01.05.2007]

77 Vgl. Mail von Manfred Hauer an inetbib@ub.uni-dortund.de vom 20.03.2007: http://www.ub.uni-dortmund.de/listen/inetbib/msg32993.html [Letzter Aufruf: 01.05.2007]

4.5 Verwendete Technologie

Dandelon.com ist der Name eines Online-Dienstes. Er basiert auf dem Zusammenspiel von drei Produkten: intelligentCAPTURE, intelligentSEARCH und IC INDEX.

4.5.1 IntelligentCAPTURE

IntelligentCAPTURE (iCapture) ist das Produktionssystem mit integriertem Scanning, OCR, PDF-Erstellung, maschineller Indexierung, Export in beliebige Bibliothekssysteme, Webspidering, Import von Verlags- und Agenturdaten, Import aus Bibliothekskatalogen sowie Datenaustausch zwischen Bibliotheken. Der gesamte Workflow vom Scannen bis zum Export ist unter Lotus Notes & Domino realisiert und speist sozusagen nebenbei auch eine Domino-Datenbank. Dandelon.com ist die Web-Schnittstelle dieser Datenbank mit einer bewusst einfachen Oberfläche, ähnlich wie bei Google, zumindest auf den ersten Blick. Eine besondere Stärke liegt bei der Bearbeitung von Büchern. Mittels Scanning, OCR und maschineller Indexierung werden vorwiegend Inhaltsverzeichnisse erschlossen. Diese Daten stehen immer auch in den jeweiligen Bibliothekskatalogen online zur Verfügung und umgekehrt schaltet dandelon.com zwecks Ausleihe automatisch auf die Bibliothekskataloge oder alternativ zum Buchhandel. Die PDF-Inhaltsverzeichnisse im Katalog des Gemeinsamen Bibliotheksverbundes sind mit den PDF-Dateien in dandelon.com identisch.

ICapture 1.0 wurde für die Vorarlberger Landesbibliothek in Bregenz entwickelt. Dort sind alle 1,5 Mio. Medien mit dem Bibliotheksverwaltungssystem ALEPH von ExLibris formal und inhaltlich erschlossen. Doch der Zugang zu den Inhaltsverzeichnissen von Büchern und anderen Publikationen fehlte zunächst noch. Deshalb wurde iCapture 3.0 unter Lotus Notes & Domino 5.08 entwickelt, um Scannen, OCR und automatische Indexierung zu integrieren. Ohne äußeres Zutun erkennt das vollständig in Lotus Notes integrierte Programm Adobe Acrobat Capture 3.0, dass eine neue Datei mittels OCR in ein PDF und einen Text für die automatische Indexierung verwandelt werden soll. Mehrere Arbeitsschritte laufen automatisch ab, und der Fortgang wird in einem Notes-View graphisch dargestellt. Die Nachbearbeitung von OCR-Fehlern mit Hilfe des Korrektureditors Quickfix ist ebenfalls implementiert. Quickfix visualisiert, bevor das PDF erzeugt wird, das OCR-Ergebnis in einem editierbaren Fenster, dergestalt dass die einzelnen Wörter nach Wahrscheinlichkeiten der Korrektheit, die das System ermittelt hat, aufsteigend gelistet werden. Ein Blick auf die erste Seite genügt in der Regel, um die Qualität zu beurteilen, das OCR-Ergebnis eventuell kurz zu editieren oder sofort zu bestätigen, womit der Workflow automatisch fortgesetzt und beendet wird.[78] Es wird immer ein zweischichtiges PDF erzeugt, d.h. im Vordergrund liegt das Image, im Hintergrund der Text. Das PDF erscheint also immer korrekt. Lesefehler beeinträchtigen nur das Ergebnis der Indexierung.

Das Scannen geschieht mittels eines Flachbettscanners von Fujitsu, d.h. die Bücher müssen wie bei einem Kopierer hochgehoben werden, um die Seiten umzublättern. Seit September 2006 gibt es eine mobile Scanstation mit feststellbaren Rollen und einer 50-Meter-Kabeltrommel sowie WLAN, die zusammen mit einem Bibliotheksmöbelhersteller entwickelt wurde. Sie ist für das Scannen zwischen oder in der Nähe von Regalen geeignet und besteht aus Scanner, Rechner und iCapture. Dazu können Stehhilfen benutzt werden. Mit Hilfe der mobilen Scanstation kann bspw. direkt im Magazin gescannt werden, sodass der Transport von Büchern aus dem Magazin und wieder zurück entfallen kann. Für den Scanvorgang kann auch

78 Vgl. Karl Rädler: „In Bibliothekskatalogen 'googlen'. Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge“. In: Bibliotheksdienst 38. Jg. (2004), H. 7/8. S. 928.

ein anderer Scanner eingesetzt werden, wenn jener bereits vorhanden ist. Das Scannen kann übersprungen werden, wenn die Dokumente bereits elektronisch vorliegen. Verschiedene Dateiformate können direkt den Weg der OCR nehmen. Die OCR-Software kann ebenfalls ausgetauscht werden, wenn z.B. spezielle OCR-Software für alte Schriften oder andere Sprachen benötigt wird.

Die komplexe Aufgabe der maschinellen Indexierung übernimmt die CAI-Engine. CAI bedeutet Computer Aided Indexing; dahinter stehen Entwicklungen des IAI (Institut für Angewandte Informationsforschung) in Saarbrücken. Mittels wörterbuchbasierter linguistischer Verfahren werden zunächst für alle Worte die Grundformen ermittelt (Morpheme). Zu diesen Spezialwörterbüchern kommen zahlreiche Regeln, Grammatiken und auch ein Thesaurus mit semantischen Relationen hinzu. Somit werden Einzelworte und typische Wortgruppen erkannt. Statistische Regeln führen zu einer Gewichtung, sodass die wichtigsten Terme ausgegeben werden. Dabei entstehen Gruppen:

- geographische Benennungen
- Personen, Unternehmen
- Branchen, Tätigkeitsfelder
- Sachdeskriptoren: Einzelwörter, die aus dem internen Thesaurus stammen, also evtl. nur sinngemäß im Ausgangstext vorhanden waren
- wichtige Worte und Phrasen aus dem Text
- maschinelle Zusammenfassung (was nur bei Volltexten, nicht aber bei Inhaltsverzeichnissen brauchbare Resultate liefert

Alle Indexierungsfelder sind in iCapture editierbar.[79]

Da iCapture auf IBM Lotus Notes & Domino basiert, ist bereits eine Standard-Suchfunktion verfügbar, nämlich die GTR (Global Text Retrieval). Sie ist eine Suchmaschine mit Stemming, Fuzzy-Suche, Feldsuche, numerischer Suche, Datumssuche, Termgewichtung und kann über mehrere Domino-Datenbanken optional gleichzeitig suchen.

4.5.2 IntelligentSEARCH

IntelligentSEARCH stellt die Retrieval-Software dar, die auf einem Vektorraum-Retrieval-Modell basiert. Sie sammelt die Daten aus intelligentCAPTURE aller Produzenten, macht sie suchbar und verbindet in die Kataloge der Bibliotheken zurück. IntelligentSEARCH unterstützt eine mehrsprachige semantische Suche. Sie nutzt die GTR-Funktionen, geht aber darüber hinaus. Suchbar sind Bücher und Artikel, vorwiegend aktuelle Titel und aus allen wissenschaftlichen Disziplinen.

Mit der Standard-Abfrage und den Standard-Einstellungen ist der Modus „more precise“ („so genau wie möglich“) aktiv, d.h. es werden Synonyme, Abkürzungen und thesaurusbasierte Übersetzungen automatisch in die Suche einbezogen (Synonymie-Expansion, z.B. Gebäude OR Haus). Bei „medium precise“ werden zusätzlich die Unterbegriffe in der Sprache der Anfrage ergänzt (Hierarchie-Expansion, z.B. Einfamilienhaus OR Wohnanlage). „More recall“ („deutlich breiter“) nutzt eine Variation der ursprünglichen Suchbegriffe durch gleichzeitige Links- und Rechtstrunkierung (*haus*) oder eine Ähnlichkeitssuche (Fuzzy-Technik). Prinzipiell finden immer eine morphologische Analyse und ein Stemming statt (Häusern OR Häuser OR Haus). Weitere Einstellungen bei der Suche betreffen die Wahl einer Vorzugsbibliothek

79 Vgl. Manfred Hauer: „iCapture 1.0 bringt Inhaltsverzeichnisse in Bibliothekssysteme und verbessert die Recherche“. In: B.I.T. Online, Heft 1, 2002. S. 50.

und die Beschränkung der Suche auf diese bzw. die Suche über alle Bibliotheken, die an dandelon.com teilnehmen.

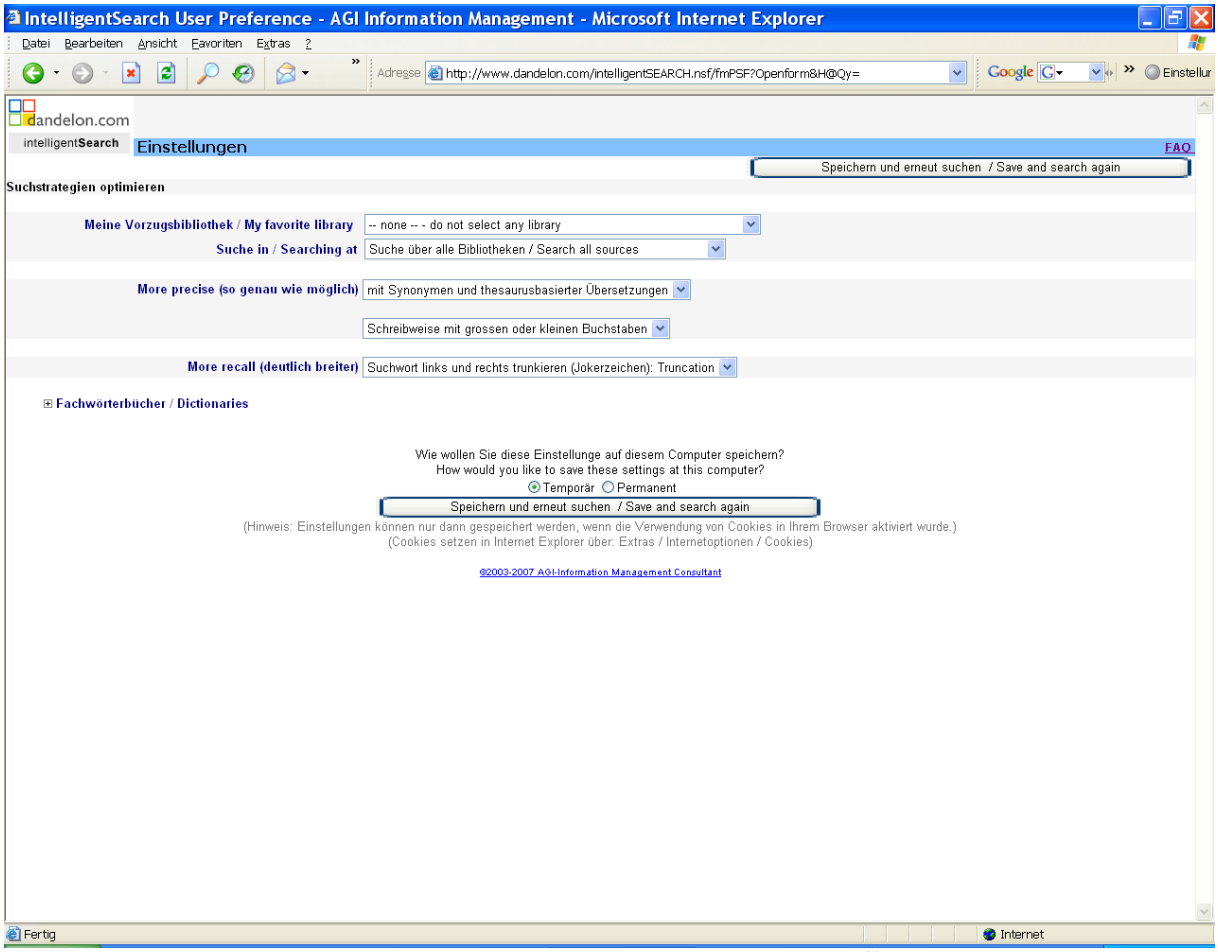

Abb. 6: Einstellungen von intelligentSEARCH[80]

Die Suchergebnisse werden nach Relevanz sortiert (Relevance Ranking). In die Berechnung der Relevanz gehen folgende Faktoren ein:

- Position des Suchworts
- Häufigkeit aller Suchworte im Dokument
- Länge des Dokuments

Was die Position betrifft, so ist ein Suchwort wichtiger, wenn es in den Feldern für den Titel (vom Autor betont), den Feldern für die intellektuelle Sacherschließung (vom Bibliothekar betont) oder unter den ersten maschinell generierten und hochgewichteten Deskriptoren (von intelligentCAPTURE generiert) zu finden ist. Je häufiger das Suchwort (und auch automatisch mit gesuchte Synonyme oder Übersetzungen) in einem Dokument vorkommt, desto höher ist seine Relevanz. Die Dokumentenlänge wirkt sich ebenfalls auf die Suchworte bzw. die Relevanz aus: sehr kurze Dokumente haben aufgrund der Relation zwischen Suchworten und den übrigen Worten immer eine hohe Relevanz. Bsp.: Ein Dokument mit 10 Worten und einem Suchwort ergibt ein Verhältnis von 1:10. Ein Dokument mit 2.000 Worten und 20 Suchworten ergibt ein Verhältnis von 1:100. Da sich die kurzen Dokumente immer vorschieben, wird hier zugunsten von langen Dokumenten manipuliert. Das Ranking wird durch eine Zahl angegeben. Werte unter 75 lassen meist auf weniger relevante Resultate schließen. Mehr als 90 wird fast nie erreicht. Für Nutzer ist das Ranking teilweise schwer nachvollziehbar, weil Index-Terme benutzt werden, die nicht sichtbar sind (z.B. aus Thesauri oder aus der

80http://www.dandelon.com/intelligentSEARCH.nsf/fmPSF?Openform&H@Qy= [Letzter Aufruf: 01.05.2007]

maschinellen Indexierung).

Die Suchergebnisse, die zunächst nach dem beschriebenen Relevance Ranking sortiert werden, können auch nach Autor und Jahr umsortiert werden. Automatisch werden bibliographische Daten und - soweit verfügbar - Klappentexte und ein Teilbereich sowohl der maschinellen als auch der intellektuellen Indexierung angezeigt.

4.5.3 IC INDEX

IC INDEX ist die Thesaurus-Entwicklungsumgebung. Hier werden Thesauri geschrieben oder importiert. Mehrere IC-Index-Datenbanken und damit Thesauri sind in intelligentSEARCH und anderen Suchapplikationen von AGI-IMC integriert. Dandelon.com hat über 1,4 Mio. Fachbegriffe aus Thesauri hinterlegt und unterstützt damit ein Cross Language Retrieval (mehrsprachige semantische Suche) in 20 Sprachen. Durch die thesaurusbasierte Übersetzung werden hierarchisch assoziierte Begriffe in die Suche eingebunden. Zugleich findet eine Fokussierung auf die akademische Fachsprache statt. Suchanfragen werden automatisch und mehrstufig optimiert. Zu den Thesauri zählen u.a. EUROVOC, AGROVOC, JuriVoc, Medical Subject Headings (MeSH), INFODATA und der Thesaurus des deutschen Bundesumweltamts. Innerhalb der Thesauri kann in einer grafischen Visualisierung navigiert werden. Die Thesauri in mehreren Sprachen und Fachgebieten, die zumeist anderen Organisationen gehören und urheberrechtlich geschützt sind, sind über den Button „Fachwörterbücher / Dictionaries“ sichtbar und nutzbar. Voreingestellt sind alle Thesauri für alle Benutzer. Die Suche wird schneller, wenn ein Benutzer nur die zu seinem Fachgebiet passenden Thesauri zuschaltet.

5 180T-Projekt

Unter dem Arbeitstitel 180T (für 180.000 Bücher) lief in Köln im Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) ein Catalogue-Enrichment-Projekt, in dessen Verlauf die Inhaltsverzeichnisse von 180.000 Büchern eingescannt, mit einer Texterkennung als Volltext aufgearbeitet und in die verschiedenen Katalogsysteme eingespeist wurden.[81]

5.1 Projektziel

Ziel des 180T-Projekts war die Erweiterung der Kataloginformation um die Inhaltsübersicht, um durch zusätzliche, bislang unerschlossene Inhaltsinformation (die Inhaltsverzeichnisse sind für die Nutzer als digitales Bild einsehbar) einen Mehrwert für den OPAC zu schaffen. Die Einbindung der Stichwörter der Inhaltsverzeichnisse in die recherchierbaren Suchbegriffe des Katalogs soll dem Benutzer ermöglichen, Texte zu finden, die allein über Titelstichwörter und Schlagwörter nicht auffindbar gewesen wären. Somit soll die Literatursuche zielgerichteter und erfolgreicher gestaltet werden; Suchende sollen auch in Aufsatzsammlungen fündig werden. Die Kataloganreicherung soll den OPAC ein wenig Amazon und Google annähern, denn die Internetnutzer sind es z.B. von Amazon gewöhnt, mehr Information zu einem Titel einsehen zu können (Cover, Rezensionen etc.). Dadurch sollen Fehlausleihen vermieden und die Nutzungsfrequenz der Titel erhöht werden; die Bibliothekskunden sollen ökonomischer vorgehen können. Das Scannen soll zugleich den überregionalen Aufgaben nützen, weil die Fernleihkunden mit zusätzlichen Informationen bei ihrer Titelauswahl unterstützt werden. Als Nebeneffekt werden vielleicht auch Dokumentenlieferaufträge für Beiträge erteilt, die sonst weiter unbeachtet in Sammelbänden im Regal verblieben wären.

5.2 Projektpartner

Das 1973 gegründete hbz ist eine zentrale Dienstleistungs- und Entwicklungseinrichtung für Bibliotheken innerhalb und außerhalb von Nordrhein-Westfalen. Wichtige Produkte des hbz sind die hbz-Verbunddatenbank, die Digitale Bibliothek DigiBib, ein Publikationssystem für Open-Access-Artikel (Digital Peer Publishing), der Dreiländerkatalog, die Deutsche Bibliotheksstatistik (DBS), Online-Fernleihe und Dokumentlieferdienste, DigiLink sowie der hbz-Medienserver. Im Verbundkatalog des hbz fließen die Daten von 246 Bibliotheken zusammen. Der hbz-Medienserver ist ein großer und stark frequentierter Datenpool, in dem den Verbundteilnehmern ca. 13 Mio. Titeldaten mit ca. 30 Mio. Exemplardaten zur Verfügung stehen. Das hbz, das die Projektkoordination übernahm, wurde vom Ministerium für Innovation, Wissenschaft, Forschung und Technologie des Landes Nordrhein-Westfalen mit ca. 220.000 € unterstützt. Das Projekt fand in Kooperation mit zwei Kölner Bibliotheken, der Universitäts- und Stadtbibliothek Köln (USB Köln) und der Deutschen Zentralbibliothek für Medizin (ZB MED) statt. Aus dem Bereich der ZB MED wurden 60.000 Monografien aus den Zugängen der vergangenen fünf Jahre bearbeitet, aus der USB Köln 120.000 Titel aus den Erwerbungen der vergangenen 15 Jahre des Fachbereiches Wirtschafts- und Sozialwissenschaften. Der unterschiedliche Zeitansatz wurde mit der unterschiedlichen Halbwertszeit der jeweiligen Fachliteratur begründet. Beide teilnehmenden Bibliotheken haben eine herausgehobene Stellung: Die ZB MED ist die zweitgrößte medizinische Fachbibliothek der Welt,

81 Die Beschreibung des Projekts folgt im Wesentlichen Astrid Großgarten: „Katalogsysteme mit Inhaltsverzeichnissen angereichert“. In: BuB 58 (2006) 10, S. 664-666.

nach Nutzerzahlen sogar die größte. Die USB Köln hat mehrere DSG-Sondersammelgebiete, umfangreiche Spezialbestände und den bedeutendsten Altbestand in NRW, weshalb sie auch überregionale und außeruniversitäre Kunden anspricht. Für das Projekt wurden Titel aus dem betriebs- und sozialwissenschaftlichen Bestand gewählt, der ca. ein Drittel des Gesamtbestandes ausmacht. Beide Fachbereiche decken neben der deutschen und angloamerikanischen Forschungsliteratur auch weitgehend die relevanten Titel aus dem gesamten europäischen Sprachraum ab. Die wissenschaftliche Relevanz der Fachliteratur sowie die ausgewählten Pilotbibliotheken belegen die Dimension des Projekts. Da von Anfang an feststand, dass die teilnehmenden Bibliotheken die zusätzliche Arbeit nicht im normalen Bibliotheksalltag bewältigen können, wurde als Dienstleister die Firma ImageWare Components aus Bonn, Hersteller von Bookeye®-Buchscannern und MyBib-Liefersystemen, beauftragt.

5.3 Projektorganisation

Für das Projekt gab es zeitliche, finanzielle und organisatorische Vorgaben. Diese sahen vor, dass:

- keine zusätzliche Hard- und Software angeschafft werden durfte,
- auf der vorhandenen Infrastruktur (Medea3-Umfeld der hbz-Verbundbibliotheken und MyBib-Server der USB Köln) aufgesetzt werden musste,
- keine erneute Mediendatenerfassung stattfinden sollte,
- der Bibliotheksbetrieb im Ablauf nicht behindert werden durfte.

Folgende Projektorganisation wurde vereinbart:

- die beteiligten Bibliotheken stellen Netzwerkverbindungen und Arbeitsräume für den Dienstleister und ermöglichen dem Dienstleistungspersonal Zutritt zu Freihandausleihe und Magazinen
- der Dienstleister stellt die technische Ausstattung
- der Dienstleister stellt einen Server für die Auftragsbearbeitung, der wiederum über eine Schnittstelle mit dem hbz-Server verbunden wird
- das hbz konfiguriert zusammen mit dem Dienstleister den Server
- in beiden Bibliotheken gibt es feste Ansprechpartner für technische Probleme, Qualitätssicherung und fachliche Fragen
- gemeinsam werden die Qualitätsstandards festgelegt und deren Einhaltung überprüft.

Das Scanning fand vor Ort in den Bibliotheken statt, wo jeweils mehrere Scanstationen aufgebaut wurden. Die Integration der Inhaltsverzeichnis-Images in den hbz-Medienserver und in die hbz-Verbunddatenbank war Aufgabe des hbz. Über den optimierten MyBib-eDoc-Server wurde die Auftragsverwaltung, -steuerung, -verfolgung und Betriebsdatenerfassung lückenlos abgewickelt. Da die Arbeitsergebnisse jederzeit von jedem Projektmitglied per Web-Zugriff auf das MyBib-System geprüft werden konnten, war die Projektgruppe stets auf dem Laufenden und konnte so die Abstimmungsbesprechungen (z.B. zur Bewertung der Arbeitsqualität oder zur Diskussion von Sonderfällen wie Kritzeleien im Inhaltsverzeichnis, mehrsprachige Verzeichnisse und solche mit Formeln oder arabischen und chinesischen Schriftzeichen, die von der Texterkennung nicht oder nur fehlerhaft erkannt wurden) auf ein Mal pro Monat begrenzen. Die Scan-Qualität sowie die OCR-Qualität wurden stichprobenhaft gemeinsam kontrolliert durch das hbz, ImageWare und die Qualitätssicherungsbeauftragten in den beteiligten Bibliotheken.

5.4 Projektverlauf

5.4.1 Pilotphase und erste Projektphase

Um Überraschungen beim Projektstart zu vermeiden, war eine Pilotphase im Juli und August 2005 vorgeschaltet. Zunächst wurden jeweils 500 Bände pro teilnehmender Bibliothek verarbeitet. Die Ergebnisse wurden von den Ansprechpartnern der Bibliotheken und des hbz hinsichtlich Scanqualität und Genauigkeit der Texterkennung überprüft. Nach der erfolgreichen Testphase ging das Projekt am 1. September 2005 in den Produktionsbetrieb und wurde planmäßig vor Weihnachten 2005 abgeschlossen.

Für die erste Projektphase wurden in der USB Köln acht und in der ZB MED vier Scanstationen aufgebaut. Die Ausrüstung pro Arbeitsplatz umfasste außer Mobiliar und Bücherwagen jeweils einen Bookeye®-GS400 mit ergonomischem Scanpad und Barcodepistole. Der Bookeye-Scanner ist ein Aufsichts- und Buchscanner, bei dem das ständige Hochheben und Umdrehen des Buches wie beim herkömmlichen Flachbettscanner, der bei dandelon.com eingesetzt wird, entfällt. Die Mitarbeiter holten die Bücher an den Scanplatz, wo jedes Buch zuerst mit der Barcodepistole registriert wurde. In den folgenden Arbeitsschritten wurden die Seiten des Inhaltsverzeichnisses gescannt und um irrelevante Informationen bereinigt. Danach wurde die Texterkennung durchgeführt. Jeder Mitarbeiter war angehalten, die Stimmigkeit des Ergebnisses zu überprüfen. Sollten Fehler übersehen worden sein, wären sie später in dem mehrstufigen Qualitätssicherungsverfahren aufgefallen. In dem Fall bekamen die Aufträge im System einen Reklamationsvermerk und erschienen auf späteren Auftragslisten zur Nachbearbeitung.

Die Mitarbeiter, die die Scanarbeiten ausführten, sahen nur die Benutzeroberflächen der verwendeten Scansoftware BCS-2®. Das recht aufwändige MyBib-System im Hintergrund, das den gesamten Workflow steuerte und die lückenlose Auftragsverfolgung von den lokalen Bibliothekssystemen zum hbz-Server erst möglich machte, war nur den jeweiligen Projektbeauftragten zugänglich. Über MyBib wurden so genannte Buchhollisten mit je 25 Titeln, die nicht ausgeliehen waren, erzeugt und an die Scanoperatoren verteilt. Diese Listen waren Auftragszettel und gaben den Mitarbeitern die zu bearbeitenden Bücher vor. Jedes Buch war über Signatur, Mediennummer und Titel auf der Liste ausgewiesen. Den Mediennummern kam eine Schlüsselrolle in dem komplexen Datengefüge zu, weil sie die eindeutige Identifizierung für ein Buch und in Form eines Barcodeetiketts auf jedem Medium aufgebracht waren. Der Barcode identifizierte das Buch gegenüber MyBib, was wiederum eine Verknüpfung zur Verbund-Identifikationsnummer (ID) des hbz herstellte.

Insgesamt wurden 180.000 Bücher aus dem Magazin geholt und wieder dorthin zurückgebracht; es kamen also keine mobilen Scanstationen wie an der SUB Hamburg im Rahmen von dandelon.com zum Einsatz. Bei 720.000 Seiten wurden an insgesamt 12 Scanstationen Scannen und Texterkennung in 20 verschiedenen Sprachen durchgeführt. Für einen Zeitraum von 90 Arbeitstagen bedeutete das 2.000 Titel sowie 7.000 Seiten pro Tag. Die erzielten Maximalwerte lagen bei 3.000 Titeln und 11.000 Seiten pro Tag.[82]

82 Die Zahlen stammen aus: Astrid Großgarten: „Catalogue Enrichment – ein Mehrwert für den Katalog am Beispiel des hbz-Projektes 180T“. Präsentation für die Jahrestagung des Arbeitskreises Bibliotheken und Informationseinrichtungen der Leibniz-Gemeinschaft in Göttingen vom 28.-29.09.2006.http://www.hbz-nrw.de/dokumentencenter/produkte/catalogue_enrichment/aktuell/vortraege/180T-Projekt_ABI_Leibniz_2006.pdf [Letzter Aufruf: 01.05.2007]

Das hbz übernahm zum Jahreswechsel 2005/2006 die gewonnenen Daten in seinen Medienserver. Die Umsetzung des durch Texterkennung generierten Volltextes erlaubte dabei eine Indexierung der Daten über die im hbz verwendete Suchmaschine, die auf der Technologie FAST beruht.

5.4.2 Zweite Projektphase

Da das Projekt erfolgreich war, wurde zum Jahresbeginn 2006 die zweite Phase eingeläutet, in der die Kataloganreicherung nicht nur retrospektiv, sondern auch für Neuzugänge betrieben wurde. Zu den teilnehmenden Bibliotheken sind die Universitäts- und Landesbibliothek Düsseldorf, die Universitätsbibliothek Paderborn sowie die Universitäts- und Landesbibliothek Bonn hinzugekommen. Die Koordination lag weiterhin beim hbz. Auch bei den neuen Teilnehmern wurden nach dem in Köln erprobten Organisationsmodell Scanarbeitsplätze installiert und Verbindungen zum zentralen MyBib-eDoc-Server geschaltet, der den kompletten Geschäftsgang überwachte. Zielvorgabe für den Zeitraum bis zum Spätsommer 2006 war es, den Workflow des Scannens von Neuerwerbungen zu evaluieren sowie die Inhaltsverzeichnisse aller Neuzugänge des Verbundbereichs im hbz-Medienserver, im Dreiländerkatalog und in der hbz-Verbunddatenbank anzureichern und nachzuweisen. Zusätzlich wurden retrospektiv Titel aus den Fachgebieten Wirtschafts- und Sozialwissenschaften, Medizin, Germanistik, Romanistik und Mathematik digitalisiert. Zum 30. August 2006 waren ca. 230.000 Titel bearbeitet. Die Ergebnisse wurden vom hbz ebenfalls in den Dreiländerkatalog und den hbz-Verbundkatalog übernommen. Im OPAC der USB Köln waren zum 25. September 2006 142.910 Inhaltsverzeichnisse eingespeist; im OPAC der ZB MED zum Oktober 2006 60.000 Retrodaten sowie 10.000 Titel aus dem Jahr 2006 inkl. Dissertationen. Dargestellt wurden die Inhaltsverzeichnisse in allen Katalogen zunächst durch den Zusatz „Inhaltsverzeichnis“ und eine URL oder durch ein Dokumentensymbol mit dem Link „Inhaltsverzeichnis“. Die folgenden Dateiformate wurden gewählt:

- OCR für den Index
- PDF für den Ausdruck
- TIFF für die Langzeitspeicherung.

Auch im Jahr 2007 setzt das hbz den begonnenen Weg des kooperativen Scannens von Neuzugängen fort. Die Tätigkeiten aus den ersten beiden Projektphasen sollen in eine Standarddienstleistung überführt werden. Evtl. gibt es auch ein Budget für das retrospektive Scannen von interessanten Beständen. 12 Bibliotheken haben sich um die Teilnahme am kooperativen Scannen beworben. Kriterien für die Auswahl sind die kritische Masse von Neuerwerbungen (10.000-12.000) und die interessanten Bestände (Sondersammelgebiete). Die ausgewählten Bibliotheken werden über hbz-Projektmittel gefördert. Die Beteiligungsmöglichkeiten für kleine Bibliotheken bestehen in der Eigenorganisation und -finanzierung der Scanleistung sowie in der Nutzung der hbz-Infrastruktur (Verbunddatenbank zur Scanbeauftragung und MyBib-System zur Scanabwicklung). Die zu erbringende Qualität wird durch das hbz vorgegeben.

5.4.3 Weitere Projektplanung

In einer weiteren Projektphase ist geplant, neben Inhaltsverzeichnissen auch Rezensionen, Klappentexte, Verlagsinformationen und Abstracts in gleicher Weise in die Bibliothekskataloge zu integrieren und nutzbar zu machen. Erste Meilensteine stellen hierbei die Kooperationen des hbz mit Springer Science + Business Media, dem Thieme Verlag und der Library of Congress dar. Ziel der Kooperation mit Springer ist eine Anreicherung der im hbz-Verbund-

bestand nachgewiesenen Springer-Titel um Inhalte wie Buchcover, Inhaltsverzeichnisse, Vorworte und Probekapitel. Der Verlag lieferte dem hbz zunächst einen Grundbestand von 22.000 Objekten der Verlagsproduktion der letzten fünf Jahre. Daran schlossen sich Ergänzungslieferungen an: Neue Titel des Verlags werden monatlich auf den Servern des hbz eingespielt. Technische Basis des Catalogue Enrichment sind das ALEPH-Verbundsystem, der hbz-Medienserver und die hbz-Suchmaschinentechnologie zur Präsentation der Daten. Bis zum Mai 2006 wurden die rund 17.000 im hbz-Verbundkatalog vorhandenen Springer-Titel in den Medienserver geladen und in der hbz-Verbunddatenbank verfügbar gemacht. Von dort aus werden sie in die lokalen Kataloge der Verbundbibliotheken und in den Dreiländerkatalog des hbz integriert. Die Daten werden soweit möglich volltextlich indexiert und können somit bei der Recherche direkt durchsucht werden. Das hbz verhandelt mit weiteren Verlagen und Datenanbietern über Kooperationen, um den Bibliotheken weiteren, über die traditionelle sachliche Erschließung hinausgehenden Mehrwert für ihre Kataloge bieten zu können.[83] Die Ausweitung der Kooperation mit Verlagen und mit anderen Verbünden sowie die Erweiterung um andere Inhaltsinformation sind Zukunftsperspektiven des hbz im Bereich Catalogue Enrichment. Um Doppelarbeit zu vermeiden, findet in der AG Kataloganreicherung eine Abstimmung mit anderen Bibliotheksverbünden statt. Ein zentrales Repository für den Nachweis aller Kataloganreicherungen der Verbünde soll auf Basis des hbz-Medienservers aufgebaut werden. Scandaten sollen untereinander ausgetauscht werden. In der Überprüfung befindet sich die Integration der hbz-Daten in Google.

Im Nachgang zum Catalogue Enrichment mussten sowohl die Mitarbeiter als auch die Nutzer informiert bzw. letztere auch geschult werden sowie die Suchmasken angepasst werden. Als „Hausaufgabe“ stehen die Überprüfung von Nutzerverhalten, Nutzererwartung und Leihverhalten und die Überprüfung der Arbeitshypothesen an.

5.5 Verwendete Technologie

5.5.1 Titelanreicherung durch Scandaten

Für die Anreicherung der Titeldaten mit Scandaten wird zum einen die oben genannte Hardware der Scanstationen benötigt; zum anderen wird Software für die Scan-Erstellung, die Scan-Verwaltung, die Titelanreicherung und die Präsentation benötigt.[84]

Bei der Scan-Erstellung sind die folgenden Komponenten beteiligt:

- Scanner mit BCS-2®-Schnittstelle/-Software
- MyBib-eDoc zur Verwaltung der Scanaufträge
- sftp im hbz zur Übernahme der Scandaten

Die Verwaltung und Langzeitspeicherung der Scandaten erfolgt über den hbz-Medienserver. Es werden die folgenden Objekte verwaltet:

- OCR-Text (Qualitätsgründe)

83 Vgl. Pressemitteilung vom 07.04.2006: „Ausweitung des Catalogue Enrichment. Hochschulbibliothekszentrum NRW kooperiert mit Springer“: http://www.hbz-nrw.de/dokumentencenter/presse/pm/springer [Letzter Aufruf: 01.05.2007]

84 Die Ausführungen zur Technik sind entnommen aus: Stephani Scholz und Hermann Kronenberg: „Catalogue Enrichment. Neue Wege der Erschließung“. Präsentation für den 95. Deutschen Bibliothekartag in Dresden vom 21.-24.03.2006. http://www.hbz-nrw.de/dokumentencenter/produkte/catalogue_enrichment/aktuell/vortraege/Medienserver_CE.pdf [Letzter Aufruf: 01.05.2007]

- PDF-Objekt für kundenfreundliche Anzeige
- TIFF-Objekt für Archivierung und ggf. erneutes Erstellen der OCR-/PDF-Objekte

Die Metadaten der mit den Titeldaten verlinkten Objekte werden über einen Replikationsmechanismus aktuell gehalten. Über Z39.50 und OAI können die Scandaten recherchiert und übernommen werden.

Was die Titelanreicherung angeht, so werden die Verbundtitel über eine Schnittstelle zwischen Medienserver und Verbundsystem automatisiert mit den Scandaten verlinkt, wenn die Scandaten über eine Verbund-Titelidentifikation verfügen. Die Metadaten des Medienservers werden durch die Verbund-Metadaten ersetzt und aktuell gehalten, wenn Objekte verlinkt werden. Bei der Verlinkung werden die OCR-Texte in die hbz-Verbunddatenbank übernommen und dort ebenfalls indexiert. Nach der Titelanreicherung werden die Links und Volltexte den Lokalsystemen mit entsprechender Schnittstelle zur Verfügung gestellt (siehe Grafik auf S. 46).

Die Präsentation betrifft die hbz-Verbunddatenbank (Katalogisierungsclient und OPAC), den Dreiländerkatalog und den hbz-Medienserver. Im OPAC werden die Inhaltsverzeichnisse dem Nutzer als PDF angezeigt. Es kann aber auch die Variante OCR-Text gewählt werden.

Der bereits erwähnte Dreiländerkatalog umfasst neben den Daten des hbz-Verbundes die Daten des Bibliotheksverbundes Bayern (BVB), des Österreichischen Bibliothekenverbundes (OBV.SG) und des Gemeinsamen Bibliotheksverbundes (GBV). In Abstimmung mit den anderen Verbünden sollen künftig auch die Katalogdaten weiterer Verbünde und Bibliotheken aus Deutschland, Österreich und der Schweiz in den Suchraum des Dreiländerkatalogs integriert werden, sodass in Zukunft ein Gesamtnachweis der Bibliotheksbestände im deutschsprachigen Raum mit einem einzigen Zugriff möglich sein soll. Durch den Einsatz von Suchmaschinentechnologie der Firma FAST können in Sekundenbruchteilen mehr als 60 Mio. Titeldaten und deren Bestandsnachweise sowie zusätzliche Informationen zu den Titeln (z.B. die oben beschriebenen Inhaltsverzeichnisse) durchsucht werden. Der Dreiländerkatalog bietet vielfältige Suchfunktionen in den bibliographischen Daten. Die Verfügbarkeit der Dokumente kann in den jeweiligen verbundeigenen Systemen ermittelt werden.

Folgende Funktionen bietet der Dreiländerkatalog:

- Intuitive Bedienung im typischen Suchmaschinen-Layout
- Sortierung der Ergebnisse nach Relevanz (Relevance Ranking)
- Ausblendung von Mehrfachtreffern
- Kategorisierung der Treffermengen (z.B. nach Themen, Jahren, Publikationstypen)
- Catalogue Enrichment: Einbindung gescannter Inhaltsverzeichnisse und Verlagsdaten
- Einbindung von Normdaten und Sacherschließungsinformationen
- Verwendung linguistischer Verfahren wie Rechtschreibvorschläge und Wortstammbildung
- Erkennung von Eigennamen und Zeitschriftentiteln
- Umkreissuche (GeoSearch = Einschränkung der Suche auf Treffer im Umkreis eines Postleitzahlengebietes)[85]

Im Dreiländerkatalog findet statt einer Metasuche eine Suche in einem gemeinsamen Index statt. Diesen Index, der für Bibliotheken relevante Inhalte zusammenführt, nennt das hbz „Suchraum“. Die Inhalte können per Suchmaschinentechnologie recherchiert, und die Ergebnisse in unterschiedliche Clients eingebunden werden:

85 Vgl. http://www.hbz-nrw.de/angebote/dlk/ [Letzter Aufruf: 01.05.2007]

> Zu Ihrer Frage nach der Weiterentwicklung: wie ich schon sagte, haben wir den Gesamtbestand der Verbunddaten für unseren hbz-Suchraum indexiert und machen die Daten so für unterschiedliche Client-Systeme recherchierbar, in zuvor nie da gewesener Schnelligkeit und mit Zusatzfunktionalitäten, die die Nutzer heutzutage von Internet-Suchmaschinen kennen und zunehmend auch bei bibliothekarischen Angeboten erwarten. [...] Was den Dreiländerkatalog betrifft, so habe ich ja schon auf die großen Vorteile von Suchmaschinentechnologie gegenüber klassischer Technologie von Bibliothekskatalogen hingewiesen: sekundenschnelle Recherche, Funktionalitäten wie Drill-Downs zur Einschränkung von Suchergebnissen – eben alles, was Nutzer von modernen Internet-Suchmaschinen kennen und erwarten – und ein gemeinsamer Index aller Datenbanken statt Metasuche.[86]

Im Suchraum werden Informationen und Dienstleistungen aus unterschiedlichsten Quellen (z.B. OPAC, Fachdatenbanken, Linksammlungen, Kataloganreicherungen) zusammengeführt und vereinheitlicht. Die so genannte Suchraumredaktion ist neben der Content-Beschaffung für die Beschreibung und Erfassung der Datenquellen, die Erarbeitung der Entscheidungsvorlagen, das Priorisieren des Konvertierungs- und Indexierungsvorgangs sowie das Monitoring der Sammlungen bis zu deren Einbindung in die Client-Anwendungen zuständig. Die Daten aus den unterschiedlichen Quellen müssen konvertiert werden, wobei folgende Aufgaben anfallen: Quellformat analysieren, Zielformat validieren, Konvertierungsvorgang überwachen, Daten sichern, Suchmaschinensoftware administrieren (erweitern oder anpassen). Zentrales Element der Zusammenführung ist der gemeinsame Index, der eine einheitliche Suche über die verschiedenen Datenquellen ermöglicht. Die Clients (z.B. Virtuelle Fachbibliotheken) können sich je nach Bedarf und nach Berechtigung aus dem Suchraum ihr Portfolio an Datenquellen zusammenstellen und diese Kollektion in Sekundenschnelle durchsuchen lassen. Von zentraler Bedeutung für den Suchraum sind die Datenschnittstellen. Sie stellen die Tore zum Suchraum dar, über die Daten integriert und abgefragt werden können. Die Content-Schnittstelle ermöglicht das Einlesen und die Aktualisierung der Daten. Das hbz-Importformat beruht auf dem W3C-Standard Resource Description Framework (RDF). Die Grundlage hierfür bildet das Dublin-Core-Metadaten-Schema zur Beschreibung von Dokumenten und anderen Objekten im Internet. Über die Anfrage-Schnittstelle können unterschiedlichste Suchanwendungen in den Daten des Suchraums recherchieren (eine Schnittstelle für alle Client-Anwendungen). Ziel des hbz ist es, den Suchraum über diese einheitlichen Schnittstellen zugänglich zu machen. Die Suchraumredaktion, die Datenkonvertierung, die Content-Schnittstelle und die Anfrage-Schnittstelle bilden zusammen mit der Suchmaschine, die im folgenden Abschnitt beschrieben wird, die Komponenten des Suchraums. Der Suchraum ist mittelfristig auch als Dienstleistung für Kunden möglich. Informationsanbieter können ihre Bestände im Suchraum indexieren lassen und über eigene Web-Clients einen schnellen und personalisierten Zugang auf diese Daten erhalten.[87]

86 "hbz - Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen: Partner der Bibliotheken und Entwickler innovativer Formen der Informationsvermittlung": 10 Fragen von Bruno Bauer an Hans Ollig, Leiter des hbz. In: GMS Medizin - Bibliothek - Information 6 (2006) 2. In elektronischer Version verfügbar unter: http://www.egms.de/en/journals/mbi/2006-6/mbi000039.shtml. [Letzter Aufruf: 01.05.2007]

87 Vgl. http://www.hbz-nrw.de/angebote/suchraum/. [Letzter Aufruf: 01.05.2007]

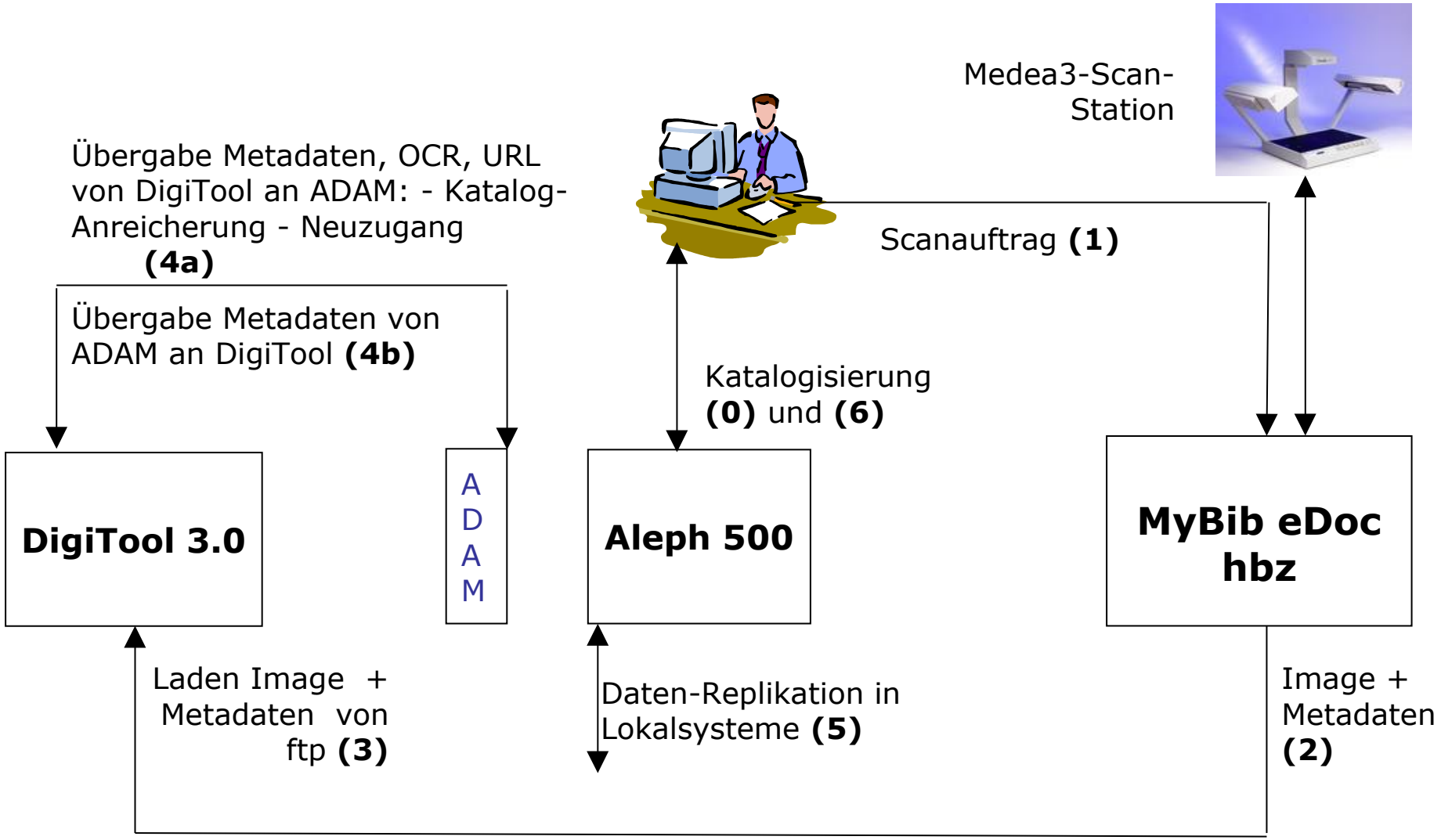

Abb. 7: Catalogue-Enrichment-Verfahren im hbz[88]

88 Stephani Scholz und Hermann Kronenberg: „Catalogue Enrichment. Neue Wege der Erschließung“, a.a.O., S. 17.

5.5.2 Suchmaschinentechnologie

Dreh- und Angelpunkt des Dreiländerkatalogs ist eine Anwendung, die im März 2005 als „hbz-Suchmaschine“ vorgestellt wurde. Die Plattform orientiert sich in Funktionalität und Layout an gängigen Web-Suchmaschinen. Die Suchresultate werden per Ranking nach Relevanz sortiert und können nach verschiedenen Kriterien angezeigt oder eingeschränkt werden, etwa Autor und Erscheinungsjahr. Die Anwendung basiert auf der Software FAST Data Search des norwegischen Unternehmens Fast Search & Transfer (FAST), das eine Ausgründung der Technisch-Naturwissenschaftlichen Universität Norwegens (NTNU) ist.[89] Anfang 2005 wurde diese Suchmaschinentechnologie in einer ersten Version auf den kompletten Datenbestand des hbz-Verbundes adaptiert. Zugriffsmöglichkeiten auf die DigiBib, die Verbundfernleihe und den WWW-Verbundkatalog wurden realisiert. Die neue Technologie zeichnet sich vor allem durch ihre Indexierungs- und Suchmöglichkeiten bei hoher Leistungsfähigkeit aus. Im Laufe des Jahres 2005 wurden die Daten aus dem BVB und dem OBV integriert. Neben notwendigen Konsolidierungsarbeiten im Bereich des Indexprofils der Suchmaschine – auch im Hinblick auf die Integration anderer Datenquellen wie z.B. des Internetportals vascoda – stand die Entwicklung einer neuen Suchoberfläche für den Dreiländerkatalog im Vordergrund. Im Oktober 2005 wurde die neue FAST-Installation durchgeführt. Die Spiegelung des Indexes erfolgt auf zwölf für den Dreiländerkatalog reservierten neuen Rechnern. Derzeit sind ca. 31 Mio. indexierte Dokumente enthalten. Die Indexierungsgeschwindigkeit liegt bei 80 Dokumenten pro Sekunde. Suchanfragen werden spätestens innerhalb einer Sekunde beantwortet, wobei mehrere hundert Mio. Dokumente durchsucht werden. Die Suchtechnologie von FAST weist laut hbz eine Reihe von Vorteilen gegenüber den traditionellen Datenbank- und Metasuchsystemen auf:

- Optimierung des Retrievals durch die Anwendung von linguistischen Verfahren, z.B. Lemmatisierung, Spracherkennung und Rechtschreibkorrektur
- Antwortzeiten im Millisekundenbereich durch das Vorhalten der Daten in einem Index, der für Endnutzerrecherchen optimiert ist
- Aufbereitung der Ergebnisliste durch ein Ranking der Treffer nach auswählbaren Kriterien (Relevanz, Erscheinungsjahr, Titel etc.)
- Analyse und Kategorisierung der gesamten Trefferliste, Einblendung von Navigatoren und Systematikbäumen zur Verfeinerung der Suche (Dynamic Drill-Down)
- Ähnlichkeitsvergleich und Suchmustererkennung[90]

Herkömmliche Rechercheanwendungen für Bibliothekskataloge und Fachdatenbanken, zum Beispiel Metasuchen in Portalen, die verschiedene Datenquellen ansprechen, funktionieren folgendermaßen: Der Nutzer gibt eine Anfrage ein. Sie wandert an die unterschiedlichen Zieldatenbanken. Von dort kommen die einzelnen Suchergebnisse zurück, werden in ein einheitliches Präsentationsformat gebracht und in einer Ergebnisliste zusammengefasst. Das Verfahren erlaubt zwar die gleichzeitige Suche über verschiedene Ressourcen, ohne dass der Anwender die Suchmaske wechseln muss, aber die Resultate können erst zusammengestellt und sortiert werden, wenn die langsamste Datenbank geantwortet hat. Läuft die Metasuche über mehrere hundert Datenbanken, kann es also entsprechend dauern, bis die komplette Liste beim Nutzer ankommt. Da Nutzer aber an die Schnelligkeit von Internetsuchmaschinen wie Google gewöhnt sind, brechen sie die herkömmlichen Suchvorgänge oft ab. FAST hingegen sammelt die Daten nicht bei einzelnen Ressourcen ein, sondern durchkämmt einen einzigen großen In-

89 FAST verkaufte seine Web Search Unit mit dem Vorzeigeprojekt der Suchmaschine Alltheweb im Februar 2003 an Overture. Overture (incl. AltaVista) wurde daraufhin im Juli 2003 von Yahoo! übernommen. FAST stellt u.a. die Suchmaschinentechnologie der Wissenschaftssuchmaschine Scirus.

90 Vgl. hbz-Jahresbericht 2005, S. 21: http://www.hbz-nrw.de/dokumentencenter/jahresberichte/hbz-jahresbericht_2005.pdf [Letzter Aufruf: 01.05.2007]

dex, der alle Verzeichniseinträge der beteiligten Bibliotheken oder Verbünde enthält. Da bei der Indexierung bereits wörterbuchbasierte linguistische Verfahren zur Anwendung kommen, muss der Nutzer keine Oder-Verknüpfung (z.B. Haus OR Häuser) eingeben, was komfortabler und ebenfalls schneller ist. Das Ranking bewertet die Anzahl der Querverweise auf den Titel in Abstracts und Fachaufsätzen. Je öfter darauf verwiesen wird, desto höher ist die Wahrscheinlichkeit, dass die Publikation im jeweiligen Fachgebiet maßgeblich ist.

Die Suchmaschinentechnologie soll künftig auch bei einem zweiten Großprojekt eingesetzt werden, an dem das hbz maßgeblich beteiligt ist: dem interdisziplinären Internetportal für wissenschaftliche Information in Deutschland, „vascoda“. Das hbz ist dabei für den technischen Betrieb und die Weiterentwicklung des Portals zuständig. Am Projekt, das vom Bundesministerium für Bildung und Forschung (BMBF) und der Deutschen Forschungsgemeinschaft (DFG) gefördert wird, beteiligen sich derzeit 40 Einrichtungen mit rund 35 Angeboten, meist nach Fachgebieten geordnete Verzeichnisse für die Online-Recherche.[91]

91 Vgl. Stefan Müller-Ivok: „Masse mit Klasse. Suchmaschinentechnologie für Bibliotheken“. München, Januar 2006. http://www.hbz-nrw.de/dokumentencenter/presse/anw/suchmaschinentechnologie [Letzter Aufruf: 01.05.2007]

6 Fazit und Perspektiven

6.1 Bewertung von dandelon.com und 180T-Projekt

Die Technologie von dandelon.com wird in verschiedenartigen Bibliotheken (Landesbibliotheken oder Universitätsbibliotheken) erfolgreich eingesetzt, entweder in einem thematisch und zeitlich klar abgegrenzten Projekt wie an der TIB Hannover oder an der SUB Hamburg, aber auch im andauernden Tagesgeschäft wie an der Vorarlberger Landesbibliothek oder an der TU Darmstadt. Beide verwenden das Produkt intelligentCAPTURE für die stetige Bearbeitung aller Neuzugänge und sind mit der Technologie zufrieden.

Es gibt aber auch Universitätsbibliotheken wie die UB Frankfurt am Main oder die UB Mainz, die die Technologie von dandelon.com getestet und sich dagegen entschieden haben. In der UB Frankfurt wurde nur die Qualität des maschinellen Indexats durch die Fachreferenten ausgewertet.[92] Für die Auswertung wurde jedem gescannten Buch ein Laufzettel beigelegt und an die Referenten verteilt. Die Referenten haben untersucht, ob bzw. inwieweit die dandelon-Deskriptoren mit einer RSWK-Erschließung übereinstimmen, inwieweit die dandelon-Deskriptoren unabhängig von RSWK als Sacherschließung geeignet sind und ob es eine Übereinstimmung der Deskriptoren mit den Titelstichwörtern gibt. Die Auswertung ergab, dass bei einem Umstieg auf dandelon starke Unterschiede im Suchvokabular auftreten würden; gleiche Sachverhalte würden nicht mehr gleich erschlossen, da die Sprache des Dokuments die Sprache der Deskriptoren bestimmt. Die Fachreferenten fanden mehrheitlich die Deskriptoren ungeeignet, sinnleer oder sogar irreführend. Wichtige Deskriptoren wie Personennamen oder geographische Benennungen fehlten. Die wichtigen Worte und Phrasen aus dem Text wurden besser als die Deskriptoren eingeschätzt. Deskriptoren sind nach einer semantischen Analyse (Abgleich mit einem Thesaurus) erzeugte Terme, wichtige Worte und Phrasen aus dem Text sind direkt aus dem Inhaltsverzeichnis entnommene Terme. Die Qualität des maschinellen Indexats wurde als zu schlecht eingestuft; außerdem wurden technische und organisatorische Probleme gesehen. Zu dem Test an der UB Frankfurt ist anzumerken, dass die Frage des Ersatzes der intellektuellen Sacherschließung durch die maschinelle Sacherschließung im Raum stand. Die Beurteilung des Indexats erfolgte durch die intellektuellen Sacherschließer, die natürlich die Gefahr sehen, dass sie sich selbst abschaffen, wenn das Ergebnis der maschinellen Erschließung gut ausfällt. Zudem geht aus den Bemerkungen hervor, dass der Test nicht frühzeitig genug angekündigt worden war, sodass die Tester nicht mit dem System vertraut waren und ihnen nicht alle Funktionalitäten klar waren. Man muss außerdem zugeben, dass gleiche Sachverhalte auch intellektuell nicht gleich erschlossen werden. Dieselbe Person erschließt zu unterschiedlichen Zeitpunkten nicht gleich; unterschiedliche Personen erschließen erst recht nicht gleich. Meiner Meinung nach geht es nicht um den Ersatz der intellektuellen durch die maschinelle Erschließung, sondern um die Kombination der beiden Verfahren, eine Ergänzung in den Fällen, die intellektuell nicht mehr zu bewältigen sind (Retroerschließung oder zu große Mengen wie z.B. im Internet). Die Testergebnisse aus Frankfurt können somit nur mit Vorsicht verwendet werden, da die Testbedingungen nicht optimal waren.

In der UB Mainz wurde ein umfangreicherer Test zum Scannen von Inhaltsverzeichnissen durchgeführt, indem die Arbeitsstation intelligentCAPTURE, dandelon.com und die Kompatibilität zum HeBIS-Verbundsystem untersucht wurden. Der Test der Arbeitsstation bezog sich auf den technischen Betrieb, den Anwenderbetrieb bzw. Workflow und die Qualität des

92 UB Frankfurt am Main: „Auswertung des Dandelon-Tests“ (unveröffentlicht, liegt der Verfasserin vor)

maschinellen Indexats; der Test von dandelon.com betraf die Endnutzerrecherche, den Dublettencheck, den Upload von Metadaten und die Technik.[93]

Die Analyse der Qualität des maschinellen Indexats gestaltete sich wie folgt: Nach Definition eines geeigneten Workflows wurden die Inhaltsverzeichnisse gescannt und für einen Satz von 20 Titeln eine Analyse der erzeugten Indexate durchgeführt. Diese teilte sich auf in eine Deskriptorenanalyse und eine Analyse wichtiger Worte und Phrasen aus dem Text. Die Qualität der inhaltlichen Beschreibung durch die Deskriptoren wurde anhand von Spezifität, Semantik und Übereinstimmung mit Titelstichwörtern bewertet. Das der Deskriptorenanalyse zugrunde liegende Verfahren wurde entsprechend auf die wichtigen Worte und Phrasen aus dem Text angewandt. Bei den untersuchten Titeln handelt es sich um Lehrbücher verschiedener chemischer Disziplinen. Die Beschreibung der Inhaltsverzeichnisse durch Deskriptoren wurde vom Tester als unbefriedigend bewertet, weil die Deskriptorrelevanz durchschnittlich nur bei 50% lag und die inhaltliche Abdeckung ungenügend war, da relevante Deskriptoren gefehlt haben. Die Anzahl der Deskriptoren stand in keiner Korrelation zum Umfang des Inhaltsverzeichnisses. Die Beschreibung der Inhaltsverzeichnisse durch wichtige Worte und Phrasen hingegen schnitt deutlich besser ab: hier lag die Relevanz durchschnittlich bei 90%. Die Anzahl der Terme nahm mit dem Umfang des Inhaltsverzeichnisses zu.

Da nach aktueller Handhabung die Indexate, deren Qualität als unbefriedigend bewertet wurde, nur in kleiner Auswahl in die Verbunddatenbank bzw. den OPAC eingebracht werden können (max. 15 Indexterme), wurde kein Mehrwert für den OPAC gesehen. Die Recherche in dandelon.com hätte den Nutzern zusätzlich vermittelt werden müssen; zudem entspricht die Oberfläche nicht dem Bibliotheksstandard.

Positiv hervorgehoben wurde die einfache Bedienbarkeit der Scanstation und die Qualität der Scans bzw. PDFs. Ein gravierendes Problem war der nicht funktionierende Dublettencheck. Darüber hinaus wurden Dokumentation, Kommunikationswege mit dem Anbieter, Wartung und ständige Verfügbarkeit (eher schwache Kooperation mit dem GBV) als problematisch eingestuft. Weiterhin als nachteilig gesehen wurden die notwendige Investition in Hardware, die zusätzlichen Kosten für zusätzliche Fachthesauri und die „alte“ Technologie des IAI, die für zusammenhängende Texte und nicht für Inhaltsverzeichnisse entwickelt worden war. Die Einsatzmöglichkeiten der Technologie von dandelon.com in der UB Mainz hätten sich auf das kostengünstige Scannen des Neuzugangs oder Teilen des Neuzugangs sowie die Integration der PDF-Links in den OPAC zwecks Sichtung der Rechercheergebnisse auf Relevanz beschränkt. Die Integration der Indexate in den OPAC war aufgrund der Qualität und der Platzbeschränkungen nicht vorgesehen. Die Suche in den eingescannten Inhaltsverzeichnissen und somit eine parallele Recherche im OPAC und im Volltext der Inhaltsverzeichnisse wäre durch die Einrichtung des Darmstädter Query-Brokers für Mainz technisch realisierbar gewesen. Das hätte meines Erachtens zu einem nicht zu unterschätzenden Mehrwert für den Nutzer sowie zu einer Effizienzsteigerung der Ausleihe verbunden mit einer Kostensenkung der Magazinvorgänge geführt. Eine Bibliothek kann sich natürlich gegen die eine und für eine andere Technologie entscheiden; das Hauptinteresse des Nutzers besteht darin, dass es überhaupt eine Kataloganreicherung gibt. Insgesamt war die UB Mainz jedenfalls nicht von dandelon.com für den laufenden Betrieb überzeugt und favorisiert ein Modell wie das des hbz.

93 UB Mainz: Abschlussdokumentation zum Test „Scannen von Inhaltsverzeichnissen“ (unveröffentlicht, liegt der Verfasserin vor)

Die Vorteile des hbz-Modells liegen darin, dass der einzelnen Bibliothek keine Kosten für zusätzliche Hardware entstehen. Die einzigen Kosten, die entstehen können, sind Arbeitskosten, wobei das Scan-Personal aus Fördermitteln finanziert wird. Die benötigte Technik ist schon im Haus; den Rest macht der Verbund, der bekanntermaßen ein sehr leistungsstarker Verbund ist. Das hbz operiert mit Größenordnungen, an die dandelon.com nicht heran reichen kann: der Dreiländerkatalog beinhaltet 60 Mio. Titel, das Suchportal dandelon.com gerade 200.000 Titel. Auch die Retro-Scanleistung von 180.000 Büchern in vier Monaten an zwölf Scanstationen ist in keinem dandelon-Teilprojekt erreicht worden; was jedoch Neuerwerbungsscans betrifft, liegen die Zahlen in Bregenz und in Köln nicht weit auseinander.

6.2 Perspektiven

Die Verbesserung der thematischen Literaturrecherche durch Kataloganreicherung, die zu einer Verbesserung der Datengrundlage des Bibliothekskatalogs führt, ist auch weiterhin eine wichtige Aufgabe der Zukunft. Sie kann nur in Kooperation zwischen den Bibliotheken bzw. den Bibliotheksverbünden bewältigt werden, indem beispielsweise Scandaten untereinander ausgetauscht werden. Die Anreicherung erfolgt zum einen durch Einscannen und Indexieren von Information, die direkt aus dem produzierten Text entnommen wird (Inhaltsverzeichnisse, Register, Cover oder Klappentexte), zum anderen durch die Verknüpfung mit Information die Rezeption des Textes betreffend (verhaltensbasierte und explizite Recommenderdienste). Die Information aus dem Text selbst ist eher sachlich und objektiv, wobei Cover und Klappentexte des Verlags in jedem Fall auch als Werbung einzustufen sind. Explizite Recommenderdienste hingegen sind subjektiv und bringen somit wertende Information in den Katalog ein. Dies erfordert ein Umdenken bei den Bibliothekaren, die bisher die Reinheit des Katalogs durch Sachinformation anstrebten. Recommendersysteme werden zurzeit noch wenig im Bibliotheksumfeld eingesetzt. Dies wird sich in Zukunft mit Sicherheit ändern, da sie eine kostengünstige Möglichkeit der Kataloganreicherung darstellen, weil sie zum größten Teil automatisiert ablaufen. Durch Online-Rezensionen und Kundenbewertungen findet zugleich eine zunehmende Verlagerung der Anreicherung auf die Nutzer bzw. Kunden der Bibliothek statt, was ebenfalls für die Bibliothek kostengünstig ist. Vorstellbar ist eine weitere Ausdehnung der Kundenaktivität im Hinblick auf Kommunikationsmöglichkeiten wie Foren oder Blogs.

Der Bibliotheks-OPAC wird sich immer mehr zu einem (Wissenschafts-)Portal weiterentwickeln. Nach Rösch stellt die Möglichkeit der Kommunikation und Kollaboration der Nutzer eine unabdingbare Funktionalität des Portals dar. Ein weiteres konstitutives Merkmal, das Internetportale von bloßen Suchmaschinen, Webkatalogen und anderen Diensten unterscheidet, liegt in der Personalisierungsoption[94]. „Mein Konto“ ist schon seit längerem im OPAC zu finden; nun halten (z.B. im OPAC der Universitätsbibliothek Karlsruhe) auch „Meine Rezensionen“ und „Meine Favoriten“ Einzug. Diese Entwicklung werden in Zukunft viele OPACs nehmen.

Die Erfordernis des bibliothekarischen Umdenkens gilt auch für die Indexierung: Das bisherige Qualitätsziel des Katalogs bestand darin, dass die Angaben im Katalog nur auf Autopsie beruhten und alles richtig sein sollte. Schlechtere, maschinell erzeugte Indexterme stehen dazu im Gegensatz. Die nicht mehr durch Autopsie zu bewältigende Literaturproduktion in

94 Vgl. Hermann Rösch: „Funktionalität und Typologie von Portalen – Infrastruktur für E-Commerce, Wissensmanagement und wissenschaftliche Kommunikation“. In: Ralph Schmidt (Hrsg.): *Information Research & Content Management. Orientierung, Ordnung und Organisation im Wissensmarkt*. Frankfurt am Main: DGI, 2001. S. 149-150.

der Wissenschaft und die Informationsflut im Internet erfordern ein Umdenken bezüglich der Indexierung: intellektuelle Indexierung durch Bibliothekare soll nicht verdrängt und ersetzt, sondern durch computergestützte und maschinelle Indexierung ergänzt und unterstützt werden. Internetquellen z.B. können aus Zeit-und Kostengründen nur automatisiert durch Linklisten erschlossen werden. Auch hier ist eine Kollaboration der Nutzer vorstellbar, indem die Nutzergemeinschaft Anteil an der Indexierung hat.

Die Anreicherung des Katalogs bezieht sich zunächst nur auf die Datengrundlage. Aber auch das Retrieval muss weiterhin verbessert werden. Internet und Internetsuchmaschinen haben einen riesigen Impuls gegeben, den Bibliotheken mit ihren klassischen OPACs aufnehmen und verarbeiten müssen. Der klassische OPAC, der zurzeit angereichert wird, wird sich weiterentwickeln müssen, wenn er nicht aussterben will, und zwar zu einem Instrument zur Erkundung von Wissen, Entdeckung neuer Wissenszusammenhänge sowie zugehöriger Dokumente mit interaktiven Navigations- und Rechercheelementen. Dies kann u.a. durch die Integration von Suchmaschinentechnologie und Navigationshilfen (z.B. grafischer Art wie Wissensnetze) in den Katalog erreicht werden. Der Katalog muss sich als wissenschaftliches Findmittel gegenüber Internetsuchmaschinen wie Google positionieren, indem er seine Vorteile herausstellt. Dazu zählen u.a. Qualität, langfristige Verfügbarkeit, Abdeckung und Ordnung. Dies muss vor allem den Nutzern klar gemacht werden, die denken, dass Internet und Internetsuchmaschinen Bibliotheken und Bibliothekskataloge ersetzen können. Aus diesem Grund müssen sowohl Retrieval als auch Layout modernisiert werden, um mit beliebten und viel genutzten Websites von Google, Amazon oder Ebay (bei letzteren handelt es sich ebenfalls um datenbankbasierte Verzeichnisse, also Kataloge) mithalten zu können. Es muss natürlich geprüft werden, was wünschenswert, machbar und realistisch ist; in jedem Fall muss die Nutzerorientierung die größte Rolle spielen. Den Umbau des klassischen OPAC zur Suchmaschine bzw. zum Portal kann sich weder eine einzelne Bibliothek noch ein einzelner kleiner Verbund leisten.

Hans Ollig, Leiter des hbz, legt deshalb den Schwerpunkt auf Kooperation und Konzentration, um auch in Zukunft bestehen zu können:

> Ich gehe davon aus, dass sich die Situation für Bibliotheken in den nächsten Jahren stark verändern wird. Möglicherweise wird sich die Zahl der Hochschulen insgesamt reduzieren; die Anforderungen an die Leistungen der Bibliotheken steigen. Dabei benötigen sie starke Partner. Die Verbundzentralen müssen im Zuge von Hochschulfreiheit, Kosten-Leistungs-Rechnung, sinkenden Personalzahlen, geplanter Evaluierung durch die Kultusministerkonferenz und steigenden Ansprüchen der Bibliotheken sicherstellen, dass sie diese Partner sein und die Anforderungen erfüllen können. Ob es in 2020 noch eine Verbundlandschaft, wie sie sich heute darstellt, geben wird, müssen wir abwarten. Es ist aber aus meiner Sicht wahrscheinlich, dass eine Konzentration von einzelnen Dienstleistungen bei den verschiedenen Verbundzentralen stattfinden wird, denn eine Mehrfacherfüllung derselben Aufgaben, wie sie derzeit zum Teil noch stattfindet, ist weder zweckmäßig noch ökonomisch sinnvoll.[95]

Der OPAC der Zukunft wird sich deshalb immer mehr von seiner ursprünglichen Funktion als Verzeichnis des Bestandes einer Bibliothek lösen und sich zu einem Meta-OPAC, einem virtuellen Katalog entwickeln. Der OPAC der Universitätsbibliothek Mannheim z.B. ist zwar bereits durch das Internet überall zugänglich, aber er ist doch immer noch an den Bestand der UB Mannheim, an den Ort Mannheim, an die URL der UB Mannheim gebunden. Der Nutzer

95 "hbz - Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen: Partner der Bibliotheken und Entwickler innovativer Formen der Informationsvermittlung": 10 Fragen von Bruno Bauer an Hans Ollig, Leiter des hbz, a.a.O.

muss dieses Vorwissen des Einstiegs haben und diese Vorauswahl treffen. Im Zuge von Internet, virtuellen Hochschulen, Home Office, (elektronischen) Lieferdiensten etc. ist dem Nutzer der Ort aber letztendlich gleichgültig. Für ihn zählt nur die Funktion. So wie den Käufer bei Amazon nicht interessiert, wo Amazon eigentlich sitzt und wie das Buch seinen Weg zu ihm findet, so interessiert den Informationssuchenden auch nicht, aus welcher UB die Information stammt; er will sie nur möglichst schnell an seinem PC-Arbeitsplatz haben. Im Rahmen einer Recherche zu einem bestimmten Thema will er nicht nur die relevante Literatur im Bestand der UB Mannheim finden, sondern die gesamte relevante Literatur. Um hier Abhilfe zu schaffen, werden bereits im Internet verschiedene Zusammenführungen angeboten. Aber die Vielfalt der konkurrierenden Angebote und Projekte (OPACs, virtuelle Kataloge, Virtuelle Fachbibliotheken, digitale Bibliotheken, Portale), die natürlich auch durch die historische Entwicklung und die föderale Struktur der Bundesrepublik Deutschland bedingt ist, ist für den Nutzer der Gegenwart leider nur schwer zu durchschauen.

7 Abbildungsverzeichnis

8 Literaturverzeichnis

Zu Kapitel 1

Ewert, Gisela/Umstätter, Walther: *Lehrbuch der Bibliotheksverwaltung*. Stuttgart: Hiersemann, 1997.

Gaus, Wilhelm: *Dokumentations- und Ordnungslehre*, 5. Aufl. Berlin und Heidelberg: Springer, 2005.

Knorz, Gerhard: „Information Retrieval-Anwendungen". In: M.G. Zilahi-Szabo (Hrsg.): *Kleines Lexikon der Informatik und Wirtschaftsinformatik*. München: Oldenbourg, 1995. S. 244-248.

Lohmann, Hartmut: *KASCADE: Dokumentanreicherung und automatische Inhaltserschließung.* Düsseldorf: Universitäts- und Landesbibliothek, 2000.

Nohr, Holger: *Grundlagen der automatischen Indexierung*, 3. Aufl. Berlin: Logos, 2005.

Salton, G./McGill, M.J.: *Information Retrieval – Grundlegendes für Informationswissenschaftler*. Hamburg: McGraw-Hill, 1987.

Schneider, Uwe/Werner, Dieter (Hrsg.): *Taschenbuch der Informatik*, 5. Aufl. München und Wien: Carl Hanser Verlag, 2004.

Im Internet:

Universität des Saarlandes, Fachrichtung Informationswissenschaft:
http://www.uni-saarland.de/fak5/fr56/
[Letzter Aufruf: 01.05.2007]

Wikipedia, die freie Enzyklopädie: http://de.wikipedia.org/wiki/Bibliothekswissenschaft
[Letzter Aufruf: 01.05.2007]

Zu Kapitel 2

Bertram, Jutta: *Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente*. Würzburg: Ergon, 2005.

DIN 31623-1: „Indexierung zur inhaltlichen Erschließung von Dokumenten". In: *Publikation und Dokumentation 2*, 3. Aufl. Berlin: Beuth, 1989. S. 275-279.

Fuhr, Norbert: „Theorie des Information Retrieval I: Modelle". In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 207-214.

Gaus, Wilhelm: *Dokumentations- und Ordnungslehre*, 5. Aufl. Berlin und Heidelberg: Springer, 2005.

Haller, Klaus/Fabian, Claudia: „Bestandserschließung“. In: Rudolf Frankenberger und Klaus Haller (Hrsg.): *Die moderne Bibliothek*. München: Saur, 2004. S. 222-261.

Knorz, Gerhard: „Informationsaufbereitung II: Indexieren“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 179-188.

Kuhlen, Rainer/Seeger, Thomas/Strauch, Dietmar (Hrsg.): *Grundlagen der praktischen Information und Dokumentation.* Band 2: Glossar, 5. Aufl. München: Saur, 2004.

Kuhlen, Rainer/Seeger, Thomas/Strauch, Dietmar (Hrsg.): *Grundlagen der praktischen Information und Dokumentation.* Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004.

Mittelbach, Jens/Probst, Michaela: *Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung. Strategien für das Bibliothekssystem der Freien Universität Berlin*. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2006 (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 183).

Nohr, Holger: *Grundlagen der automatischen Indexierung*, 3. Aufl. Berlin: Logos, 2005.

Panyr, Jiri: *Automatische Klassifikation und Information Retrieval*. Tübingen: Niemeyer, 1986.

Reimer, Ulrich: „Verfahren der automatischen Indexierung“. In: Rainer Kuhlen (Hrsg.): *Experimentelles und praktisches Information Retrieval*. Konstanz: Universitätsverlag, 1992. S. 171-194.

Sachse, Elisabeth/Liebig, Martina/Gödert, Winfried: *Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt*. Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft Bd. 14. Köln: Fachhochschule, Fachbereich Bibliotheks- und Informationswesen, 1998.

Schneider, Uwe/Werner, Dieter (Hrsg.): *Taschenbuch der Informatik*, 5. Aufl. München und Wien: Carl Hanser Verlag, 2004.

Womser-Hacker, Christa: „Theorie des Information Retrieval III: Evaluierung“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 227-235.

Im Internet:

Skriptum zur Vorlesung im SS 06 „Information Retrieval“ von Norbert Fuhr:
http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf
[Letzter Aufruf: 01.05.2007]

Wikipedia, die freie Enzyklopädie: http://de.wikipedia.org/wiki/Information_Retrieval
[Letzter Aufruf: 01.05.2007]

Zu Kapitel 3

Bertram, Jutta: *Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente*. Würzburg: Ergon, 2005.

Kind, Joachim: „Praxis des Information Retrieval“. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. Aufl. München: Saur, 2004. S. 389-398.

Lohmann, Hartmut: *KASCADE: Dokumentanreicherung und automatische Inhaltserschließung.* Düsseldorf: Universitäts- und Landesbibliothek, 2000.

Nohr, Holger: *Grundlagen der automatischen Indexierung*, 3. Aufl. Berlin: Logos, 2005.

„OSIRIS – Osnabrück Intelligent Research Information System“. In: ABI-Technik 20, 2000, Nr. 1. S. 89-90.

Recker, Ingrid/Ronthaler, Marc/Zillmann, Hartmut: „OSIRIS – ein Hyperbase Front End System für OPACs“. In: Bibliotheksdienst 30. Jg. (1996), H. 5. S. 833-848.

Reimer, Ulrich: „Verfahren der automatischen Indexierung“. In: Rainer Kuhlen (Hrsg.): *Experimentelles und praktisches Information Retrieval*. Konstanz: Universitätsverlag, 1992. S. 171-194.

Ronthaler, Marc/Zillmann, Hartmut: „Literaturrecherche mit OSIRIS. Ein Test der OSIRIS-Retrievalkomponente“. In: Bibliotheksdienst 32. Jg. (1998), H. 7. S. 1203-1212.

Schulungsunterlagen der ehemaligen Transmedia Projekt- und Verlagsges. mbH, Mannheim.

Im Internet:

Informationen zum Projekt KASCADE:
http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/kascade
[Letzter Aufruf: 01.05.2007]

Informationen zum Projekt MILOS:
http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/milos
[Letzter Aufruf: 01.05.2007]

Zu Kapitel 4

Hauer, Manfred: „iCapture 1.0 bringt Inhaltsverzeichnisse in Bibliothekssysteme und verbessert die Recherche“. In: B.I.T. Online, Heft 1, 2002. S. 49-51.

Rädler, Karl: „In Bibliothekskatalogen 'googlen'. Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge“. In: Bibliotheksdienst 38. Jg. (2004), H. 7/8. S. 927-939.

Im Internet:

Hauer, Manfred: „Strukturierung, Erschließung und Präsentation von Nachrichtentexten“. In: Wissensmanagement: Strategie, Prozesse, Communities (Tagungsband 2002), S. 101-108.
http://www.agi-imc.de/internet.nsf/RahmenDeutsch?OpenFrameSet
[Letzter Aufruf: 01.05.2007]

Informationen zu dandelon.com (FAQ):
http://www.agi-imc.de/icapture/FAQ_in_iS.nsf
[Letzter Aufruf: 01.05.2007]

Mail von Manfred Hauer an inetbib@ub.uni-dortund.de vom 27.01.2007 http://www.ub.uni-dortmund.de/listen/inetbib/msg32251.html
[Letzter Aufruf: 01.05.2007]

Mail von Manfred Hauer an inetbib@ub.uni-dortund.de vom 20.03.2007: http://www.ub.uni-dortmund.de/listen/inetbib/msg32993.html
[Letzter Aufruf: 01.05.2007]

Suchportal dandelon.com:
http://www.dandelon.com/intelligentSEARCH.nsf/fmQSF?OpenForm
[Letzter Aufruf: 01.05.2007]

Versandbuchhandlung Missing Link:
http://www.missing-link.de
[Letzter Aufruf: 01.05.2007]

Zu Kapitel 5

Großgarten, Astrid: „Katalogsysteme mit Inhaltsverzeichnissen angereichert“. In: BuB 58 (2006) 10, S. 664-666.

Im Internet:

Großgarten, Astrid: „Catalogue Enrichment – ein Mehrwert für den Katalog am Beispiel des hbz-Projektes 180T“. Präsentation für die Jahrestagung des Arbeitskreises Bibliotheken und Informationseinrichtungen der Leibniz-Gemeinschaft in Göttingen vom 28.-29.09.2006.

http://www.hbz-nrw.de/dokumentencenter/produkte/catalogue_enrichment/aktuell/vortraege/180T-Projekt_ABI_Leibniz_2006.pdf
[Letzter Aufruf: 01.05.2007]

"hbz - Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen: Partner der Bibliotheken und Entwickler innovativer Formen der Informationsvermittlung": 10 Fragen von Bruno Bauer an Hans Ollig, Leiter des hbz. In: GMS Medizin - Bibliothek - Information 6 (2006) 2. In elektronischer Version verfügbar unter:
http://www.egms.de/en/journals/mbi/2006-6/mbi000039.shtml
[Letzter Aufruf: 01.05.2007]

hbz-Jahresbericht 2005:
http://www.hbz-nrw.de/dokumentencenter/jahresberichte/hbz-jahresbericht_2005.pdf
[Letzter Aufruf: 01.05.2007]

hbz-Pressemitteilung vom 07.04.2006: „Ausweitung des Catalogue Enrichment. Hochschulbibliothekszentrum NRW kooperiert mit Springer“:
http://www.hbz-nrw.de/dokumentencenter/presse/pm/springer
[Letzter Aufruf: 01.05.2007]

Informationen zum Dreiländerkatalog:
http://www.hbz-nrw.de/angebote/dlk/
[Letzter Aufruf: 01.05.2007]

Informationen zum Projekt 180T:
http://www.hbz-nrw.de/angebote/catalogue_enrichment/scanaktivitaeten
[Letzter Aufruf: 01.05.2007]

Informationen zum Suchraum:
http://www.hbz-nrw.de/angebote/suchraum/
[Letzter Aufruf: 01.05.2007]

Müller-Ivok, Stefan: „Masse mit Klasse. Suchmaschinentechnologie für Bibliotheken“. München, Januar 2006.
http://www.hbz-nrw.de/dokumentencenter/presse/anw/suchmaschinentechnologie
[Letzter Aufruf: 01.05.2007]

Scholz, Stephani/Kronenberg, Hermann: „Catalogue Enrichment. Neue Wege der Erschließung“. Präsentation für den 95. Deutschen Bibliothekartag in Dresden vom 21.-24.03.2006.
http://www.hbz-nrw.de/dokumentencenter/produkte/catalogue_enrichment/aktuell/vortraege/Medienserver_CE.pdf
[Letzter Aufruf: 01.05.2007]

Zu Kapitel 6

Rösch, Hermann: „Funktionalität und Typologie von Portalen – Infrastruktur für E-Commerce, Wissensmanagement und wissenschaftliche Kommunikation“. In: Ralph Schmidt

(Hrsg.): *Information Research & Content Management. Orientierung, Ordnung und Organisation im Wissensmarkt.* Frankfurt am Main: DGI, 2001. S. 142-154.

UB Frankfurt am Main: „Auswertung des Dandelon-Tests“ (unveröffentlicht, liegt der Verfasserin vor)

UB Mainz: Abschlussdokumentation zum Test „Scannen von Inhaltsverzeichnissen“ (unveröffentlicht, liegt der Verfasserin vor)

Im Internet:

"hbz - Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen: Partner der Bibliotheken und Entwickler innovativer Formen der Informationsvermittlung": 10 Fragen von Bruno Bauer an Hans Ollig, Leiter des hbz. In: GMS Medizin - Bibliothek - Information 6 (2006) 2. In elektronischer Version verfügbar unter:
http://www.egms.de/en/journals/mbi/2006-6/mbi000039.shtml
[Letzter Aufruf: 01.05.2007]

Erklärung

Hiermit erkläre ich, dass ich diese schriftliche Hausarbeit eigenständig und ohne fremde Hilfe angefertigt habe. Alle Zitate sind als solche kenntlich gemacht.

....

(Ort, Datum) (Unterschrift)